Cross-Species Characterisation of Alternative Splicing Patterns

Tovar Corona, Jaime

Award date:
2014

Awarding institution:
University of Bath

Cross-species characterisation of alternative splicing patterns

Jaime Manuel Tovar Corona

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology and Biochemistry

December 2013

COPYRIGHT

Table of Contents

Acknowledgements

I first would like to thank my supervisor Dr. Araxi Urrutia for her support, advice and patience. Being part of her laboratory has been a tremendous opportunity and I'm profoundly thankful for everything I have learned while working under her supervision. I would also like to thank other past, present and visiting academic staff in the department, Prof. Laurence Hurst, Dr. Matthew Wills, Dr. Nicholas Priest, Dr. Humberto Gutierrez, Dr. Paula Kover, Prof. Tamás Székely and Dr. Steve Dorus; all of them have undoubtedly enriched my experience during my research studies with their guidance, support and helpful discussion.

This work would have been impossible without the support from the CONACyT postgraduate scholarship program.

Next, I thank past and present members of Araxi Urrutia's lab. Including my dear friend and research comrade Dr. Chen Lu, Atahualpa Castillo and Stephen Bush who greatly contributed to my research work and Jimena Monzón, Nina Ockendon, Wang Wei, Adrián Arellano and Marina Angelopoulou.

I cannot forget to thank other friends who I met in the University Dr. Laura López, Dr. Claudia Weber, Dr. Araceli Argüelles and Dr. Dr Rene van Dijk. I have to thank them for their support and their patience during my grumpy days. Also need to express my gratitude for the Mexican community in Bath, fencing and gym buddies, and the people I have met during my "learning Russian language" adventures. It will be impossible to name each fantastic person I have met in those groups.

I'm very grateful to my old friends back home, Dinorah Rivera, Omar Bayardi and Javier Cordoba who have remained close friends for many years. Special thanks for Victoria Tishechkina, who, for many years has been very close and dear to me.

Finally I would like to express my endless gratitude to my mother and the rest of my fantastic family who are certainly the foundation of everything I have achieved so far.

Contributions

All work presented in the main body of this thesis is my own with the following exceptions:

1. Functional characterisation of alternatively spliced genes in chapters 2-4 was carried out in collaboration with Atahualpa Castillo-Morales.

2. Illumina RNA-seq confirmation of alternative splicing sites in chapter 4 was carried out in collaboration with Dr Lu Chen

3. Chapter 5 was jointly written by Dr Lu Chen and me. Full citation for the published version of this chapter is as follows:

   Chen L, Tovar-Corona JM and Urrutia AO. 2012. Alternative splicing: a potential source of functional innovation in the eukaryotic genome. **International Journal of Evolutionary Biology** 12: ID 596274.

In Section 1 from the Appendices, I designed the pipeline to identify alternative splicing events and characterised alternative splicing events in cancer and normal tissue transcript libraries in collaboration with Dr Lu Chen. I also contributed critical comments, edited the draft and revised manuscript.

In Sections 2 and 3 from the Appendices I contributed analyses relating to alternative splicing only. I also contributed critical comments and edited drafts and revised manuscripts.

.

Abbreviations

AS	Alternative splicing
cDNA	Complementary DNA
dsx	Doublesex gene
ES	Exon skipping
ESE	Exonic splicing enhancer
ESS	Exonic splicing silencer
EST	Expressed sequence tag
GD	Gene duplication
GFS	Gene family size
GMAP	Genomic Mapping and Alignment Program
GO	Gene ontology
IR	Intron retention
ISE	Intronic splicing enhancer
ISS	Intronic splicing silencers
mRNA	Messenger RNA
NMD	Nonsense-mediated mRNA decay
PGLS	Phylogenetic generalized least squares
pre-mRNA	Precursor messanger RNA
RT-PCR	Reverse transcriptase-polymerase chain reaction
TCF7L2	Transcription factor 7-like 2
3S	Alternative 3' splice site
3S5S	Alternative 3' and 5' splice sites
3'ss	3' splice site
5S	Alternative 5' splice site
5'ss	5' splice site

Abstract

Alternative splicing is a common post-transcriptional process in eukaryote organisms by which a single gene can produce more than one distinct transcript. First discovered in the late 1970s, alternative splicing has been the focus of intense attention after the release of the human genome draft revealed a lower than expected gene number. Almost all human protein coding genes are now known to be alternatively spliced. However, how alternative splicing in humans and other well studied model organisms compares to other less characterised taxa such as protists and fungi or what is the functional role of alternative splicing remains poorly understood. Here I analyse alternative splicing in dozens of species using millions of partial transcript sequences ESTs. By applying a transcript normalisation method I showed that alternative splicing in protists and fungi is higher than previously reported and highly variable. I further observed that in representatives of both taxa, associations with translation are overrepresented among alternatively spliced genes. However, no evidence for a relationship between alternative splicing and complex phenotypes was found. Taking human lice as a model I explored the role of alternative splicing in the evolution of phenotypic variants. I found that, despite the fact that the transcriptome profiles of head and body lice are nearly identical, there are markedly differences in alternative splicing patterns. Development related functional associations were found to be enriched among genes with body lice specific alternative splicing events but not in head lice consistent with a scenario of differential patterns of alternative splicing contributing to the phenotypic innovations as human lice adapted to life in human clothing. I further explore the functional relevance of alternative splicing and its possible role in driving genomic innovations even preceding events of gene duplication. Together the work presented show that alternative splicing is widespread among previously understudied fungi and protist species and provide insights on its role in species adaptation to novel environments in using human lice as a model.

1 Introduction

Alternative splicing (AS) is a posttranscriptional process which allows a single gene to produce multiple mRNA variants (Graveley 2001). It is a complex process which happens during gene expression, between transcription and translation where mRNA sequences are generated from nuclear precursor mRNAs (pre-mRNA).

1.1 Alternative splicing

In eukaryotic protein-coding genes, coding regions can be intervened by non-coding sequences called introns (Figure 1.1). These intronic sequences are transcribed along with the coding regions and flanking untranslated regions (UTRs) but generally they do not contribute to the protein products encoded by genes. This process involves splicing out intronic sequences from pre-mRNA and then protein coding segments (exons) are linked together to produce a mature mRNA which later is translated into a protein product (Black 2003).

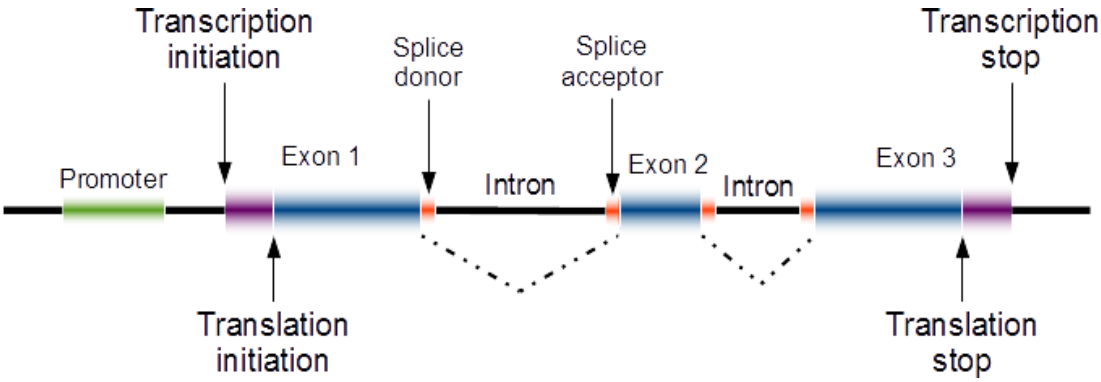


Figure 1.1. Structure of a eukaryotic gene, eukaryote genes contain one or more protein-coding regions (exons) separated by non-protein-coding regions (introns). Enhancer and promoter regions proximal to the gene regulate gene expression. From start to end the genic DNA sequence is transcribed as a unit into pre-mRNA. After the transcription of the gene, introns are excised and exons are spliced together before the transcript can be processed and the coding regions translated into a protein.

In the late 1970s splicing was first documented while observing splicing products during the lytic infection of eukaryotic cells by adenovirus 2 (Chow et al. 1977; Berget et al. 1977). These same experiments provided the first evidence for the process of alternative splicing where distinct transcripts of different lengths were generated from the same gene. Early 1980s studies found that AS is not an isolated event on cells infected by a virus (Alt et al. 1980; Early et al. 1980). These studies showed that a single endogenous gene encoded both membrane-bound and secreted antibodies through AS in mouse B cells.

Although it was once considered a rare phenomenon the evidence for ubiquity and its high incidence keeps mounting. There are examples of species of high prevalence of AS (human, 95% (Pan et al. 2008; Wang et al. 2008)) and of genes with the potential to produce a considerably large number of different transcripts through AS (*D. melanogaster*, *Dscam* gene (Schmucker et al. 2000)).

1.2 Identifying and characterising alternative splicing

Alternative splicing is the product of the interaction of a network of splice regulation factors, because of its complexity, it is not trivial to identify and understand those factors just by observing genomic properties (Barash et al. 2010; House & Lynch 2008). In fact the splicing machinery has been described as an adaptive mechanism which accommodates both local and environmental factors during the splicing process (House & Lynch 2008) making it impossible to rely on the observation of regulatory motifs in genomic data to predict AS (Barash et al. 2010).

Alternative splicing event identification primarily relies in the use of transcription products compared against each other or aligned to genomic sequences looking for signs of AS. There are several platforms which can be used to identify instances of alternative splicing on a genome wide scale. Using splice-junction microarrays, exon arrays or other microarray variations is possible to target both exons and exon junctions to test for sequence expression. Different versions of the

method exist depending on the length of the sequence which needs to be surveyed (reviewed in (Moore & Silver 2008)), but all versions have limitations because of the amount of information each microarray provides. Also, using microarrays only confirms the transcription of known sequences which implies prior knowledge of the AS sequence. A decade ago it was demonstrated it is possible to develop genome-wide surveys of AS using reverse transcriptase-polymerase chain reaction (RT-PCR) to monitor splicing of exon-exon junctions (Johnson et al. 2003). This method not only provides confirmation about the event but also provide information about the expression levels. As EST libraries vastly developed, the RT-PCR method became cost-efficient only when targeting specific genes.

Alternative splicing detection was facilitated with the possibility to sequence a large amount of mRNA sequences which could be compared to a genomic reference. This new approach, using EST libraries, completely changed the perspective about how widespread the AS phenomenon is (Artamonova & Gelfand 2007).

EST libraries have been used for a long time to identify AS events in expressed genes. Early studies in human identified the close relationship between EST library coverage and levels of AS detected (Modrek et al. 2001; Johnson et al. 2003). If the EST coverage is not taken into consideration, contradictory signals may arise about the prevalence of AS due to sampling bias (Kim et al. 2004; Brett et al. 2002). This effect was again demonstrated while comparing closely related mammal species and obtaining dissimilar prevalence of AS (E. Kim et al. 2007).

The solution is to compensate for EST coverage using a randomisation method (E. Kim et al. 2007). Averaging the AS levels detected in large number of randomly selected EST subpopulations of the same size, provides with a comparable value for AS prevalence (E. Kim et al. 2007).

With novel deep sequencing techniques transcript availability will greatly increase. RNA-seq will soon provide a clear view of diversity and expression levels for independent transcript isoforms, with a fraction of the effor. Also, with RNA-seq it will be possible to gather significant amounts of data about expression levels with

little prior knowledge about each particular isoform (Malone & Oliver 2011). But currently, there is still insufficient amounts of RNA-seq data for broad cross-species comparative studies. Recent studies in vertebrates (Barbosa-Morais et al. 2012) only include a small fraction of species with fully sequenced genomes. Challenges, like the study of alternative spliced isoforms which are lowly expressed, will still need to be addressed even after RNA-seq becomes abundant. Until RNA-seq data is available for a large number of species, EST data is to be considered a valuable source of data to study AS from a cross-species perspective.

In chapter 2, I outline the identification of alternative splicing events from partial transcript sequences using a normalisation method to correct for transcript coverage.

1.3 Alternative splicing and proteome size

It was previously thought that an organism's complexity should be reflected in the number of protein encoding genes in the organism's genome. But as the number of sequenced genomes kept growing, there was little supporting evidence for a relationship between gene number and organism complexity. While early estimates of gene number for the human genome went into six figures (Fields et al. 1994), they successively were revised downwards over the years. Gene number estimates ahead of the release of the human genome draft sequence, were in the region of 30000-40000 protein-coding genes (Lander et al. 2001). However the final release found that the human genome contains 20000-25000 genes (International Human Genome Sequencing Consortium 2004). In genome sequences of various eukaryotes, including the yeast *S. cerevisiae* (6000 genes (Goffeau et al. 1996)), the fruit fly (13600 genes (Adams 2000)) and the nematode *C. elegans* (20000 genes (Caenorhabditis elegans Sequencing Consortium 1998)), number of genes contradicted expectations, complex organisms did not show considerable larger number of genes (Hahn & Wray 2002). Therefore it was suggested the availability of more genomes will keep adding to the paradox (Hahn & Wray 2002). The mismatch

between gene number and organism complexity was coined as the G paradox (Hahn & Wray 2002), following from the previously coined C-value paradox which referred to the mismatch between genome size and complexity (from the 1971 review (Thomas 1971)). The sequencing of the mouse genome (Waterston et al. 2002) and other eukaryotic species have confirmed that gene number has remained relatively unchanged in metazoan species over the last 800 million years.

As alternative splicing provide the means to expand the transcript pool by allowing a single gene to encode more than one protein isoform (Black 2000; Graveley 2001) it soon became a strong candidate to explain the apparent missing information in the human genome (Lander et al. 2001). Using next generation sequencing techniques it is currently estimated that 95% of multi-exon genes undergo AS in the human genome (Pan et al. 2008; Wang et al. 2008). These high estimates are, in principle, consistent with AS potentially being a determining factor on proteome size.

Alternative splicing has been documented in all major taxa of eukaryotic species, further suggesting that this process may be an important player in the diversification of the transcript pool. However, whether alternative splicing can explain the mismatch between gene number and organismal complexity remains unknown. This in part is because we poorly understand how AS in human and AS in other organism relates.

Unlike many other genomic features which can be inferred from signals in genomic sequences, AS remains primarily assessed through the alignment of mRNA transcripts to the genome sequence, and then looking for variations in the coding sequence content. Although the genomes of a high number of species are now available, abundant transcript data remains scarce and patchy along vast areas of the eukaryotic tree. Even where transcript data was available at similar coverage for a large number of species, detection levels of alternative splicing are highly dependent on transcript coverage. This is because the more transcripts are available for any given gene, the greater the chances that a higher proportion of its alternative spliced transcripts will be sampled (Kim et al. 2004; Cuperlovic-Culf et al. 2006; Holste &

Ohler 2008; Yeo et al. 2004; Brett et al. 2002; Schindler et al. 2008; Roux & Robinson-Rechavi 2011; Barbazuk et al. 2008; E. Kim et al. 2007). Thus, it is not possible to assert whether alternative splicing has higher prevalence in vertebrates compared to invertebrate species or how prevalence of alternative splicing has changed over evolutionary time (Nilsen & Graveley 2010).

1.4 Prevalence of alternative splicing in eukaryotic genomes

First estimates for the number of genes with AS in the human genome were well below 50% (Mironov et al. 1999; Hanke et al. 1999; Brett et al. 2000) and then were later revised upwards (Johnson et al. 2003). Using the latest next generation sequencing techniques, more recent studies have shown that around 95% of multi-exon genes undergo AS in the human genome (Pan et al. 2008; Wang et al. 2008). Although prevalence in metazoans model species greatly vary, evidence suggests it is not at all rare (*D. melanogaster* 60% (Graveley et al. 2011) and *C. elegans* 25% (Ramani et al. 2011))

In addition to metazoans, the study of AS has been extended to include a more diverse set of eukaryote species. Individually, AS prevalence has been characterised in several eukaryote species. Results show alternative splicing is practically ubiquitous in eukaryotes. But prevalence per species has been constantly updated. Such is nothing but an expected behaviour, transcript dependant characterisation of AS are highly dependent on transcript coverage (Kan et al. 2002).

For example, in plants where AS prevalence is considered to be low, or at least lower than metazoans, the prevalence of AS for *A. thaliana* has raised constantly, from less than 2% (Zhu 2003; Zhu et al. 2003) to 61% (Marquez et al. 2012) in just 10 years. In fungi, the prevalence of AS in *A. flavus* was originally estimated to affect about 1.6% of its genes (Chang et al. 2010) the figure was later updated to 15.4% (Chang & Muddiman 2011).

Prevalence of AS has been reported in few fungi ((Chang & Muddiman 2011; Loftus et al. 2005; Hirschman et al. 2006; E. Kim et al. 2007; Zhao et al. 2013) and reviewed in (Kempken 2013)) and protist species (Iriko et al. 2009; Xiong et al. 2012; Grisdale et al. 2013). Reported prevalence is lower than that reported in plants and metazoans. But different methodologies were used and there is no correction for transcript coverage. Therefore is not possible to draw conclusions about prevalence of AS in fungi and protists from these reports.

The splicing machinery is highly conserved between plants and metazoans (Reddy 2007). But so far lower prevalence of AS has been reported in plants, for example *A. thaliana* with 61% (Marquez et al. 2012). In plants, genes seem to have followed different strategies to increase protein diversity, in plants, full genome duplication and gene neofunctionalisation is a common promoter of functional diversity (reviewed in (Moore & Purugganan 2005; Ober 2005)). Still in some cases more than half of plant genes are estimated to use AS and this proportion is expected to grow as more tissue specific AS is studied.

Metazoan is the group where we find the highest proportions of AS. In invertebrates, specially model species like *D. melanogaster* or *C. elegans*, AS has been widely documented. Even though invertebrates possess fewer and shorter introns (Deutsch & Long 1999), which are associated with a limited capacity to accommodate high proportions of AS, both species show 60% and 25% of their genes with AS (Gerstein et al. 2010; Ramani et al. 2011; Graveley et al. 2011). The proportion of genes with AS grows along the metazoan tree of life as we get closer to vertebrates (Keren et al. 2010).

Evidence has been accumulating pointing to vertebrates as the group better genomically equipped to utilise AS. Vertebrate introns have increased in length, with short ancestral introns and longer recent introns in primates (Gelfman et al. 2012; Schwartz et al. 2008) where longer introns have been associated with higher levels of AS (E. Kim et al. 2007). Cis-regulatory elements were developed from early vertebrates until the current versions in mammals, allowing longer introns (Gelfman et al. 2012). Alternatively spliced exons get shorter along the vertebrates, with the

shortest found in mammals (Gelfman et al. 2012; Sorek et al. 2004). So far, the mammal species with the highest proportion of AS is human, with 95% of multi-exon genes showing evidence of AS (Pan et al. 2008; Wang et al. 2008).

Efforts to clarify the prevalence of AS continue, and a clear example is the availability of online resources which include multiple species (reviewed in (Kim & Lee 2008; Tang et al. 2013)) and multi-species studies (McGuire et al. 2008) which now accommodate species from a wide-range of branches of the eukaryotic tree of life.

In chapter 2 and 3, I characterise alternative splicing in several protist and fungi species. I show that AS in not only ubiquitous in both protists and fungi, but also challenge previous conceptions about the prevalence of AS in this two diverse eukaryote groups.

1.5 Alternative splicing in organism adaptation

Alternative splicing events in an ever increasing list of genes have been shown to play key roles in a variety of processes in the organism. One prime example is the fruit fly *Dscam* gene which has been shown to encode multiple isoforms which play key roles in neuronal wiring (Schmucker et al. 2000; Matthews et al. 2007; Hattori et al. 2007) and immune response (Watson et al. 2005). Another textbook example is that of doublesex (dsx) gene in fruit fly, involved in the sex-determination pathway. Alternatively spliced isoforms of dsx are differentially expressed in each sex (Baker & Wolfner 1988) regulating somatic sex differentiation. The AS dependent sex specification pathway has been found to be highly conserved in insect species (Salz 2011). Together with dsx, antennapedia, grainyhead and ultrabithorax are also alternatively spliced transcription factors, the isoforms they encode are known to be related to development (antannapedia (Bermingham & Scott 1988)), to alter the function of neuroblasts (grainyhead (Uv et al. 1997)) and to regulate splicing in the presence of long introns (ultrabithorax (Hatton et al. 1998)).

In *drosophila* a single gene (Mhc) can produce close to 500 distinct isoforms through AS. Isoforms from these gene are differentially expressed in different kinds of muscles (Hastings & Emerson 1991), and are known to regulate the contractile properties of the type of muscle where they were expressed (Swank et al. 2002).

Not always is it possible to identify the functions of isoforms product of AS. But there are examples of evolutionarily conserved AS, suggesting they are functional. CadN gene is differentially spliced in embryonic mesoderm and neurons in fruit fly and in beetles (Hsu et al. 2009). In the case of unicellular fungus *C. neoformans* function of alternatively spliced genes was inferred by studying the conservation of AS while comparing with a plant species and four metazoans. Functions associated with multicellularity were not conserved across species. But molecular functions (like protein kinase activity, RNA binding and calcium ion binding) were highly conserved across all species (Irimia et al. 2007).

There are important aspects for specificity in AS like species specific, development stage specific, sex specific and tissue specific AS. Tissue specific AS events are closely associated with specific tissue functions (Taliaferro et al. 2011; Wang & Burge 2008). Nevertheless, what we know about the functional roles of AS is probably just a fraction of what is happening and further work is necessary.

In chapter 4, I investigate the role of alternative splicing in the adaptations of human lice to a lifestyle which requires for the host to use cloths. I show that even though transcriptional profiles of both lice are similar, is possible to find AS events exclusively associated with either head or body lice. When comparing the functional characterisations, I show genes with AS events exclusively associate with body lice, are associated with different functions. In the same chapter I provide extensive discussion for the effects each of these enriched functions may have on the body lice phenotype and how this could be providing the tools to exploit a new ecological niche.

1.6 Functional relevance of alternative splicing

Alternative splicing research has changed the way we understand the relationship between genes, their transcription, the protein products and their functions. The model of one gene encoding one protein has been challenged by mounting evidence pointing toward genes with the capacity to encode multiple transcripts and therefore multiple protein products. Consequently alternatively spliced genes have the potential to fulfil multiple functions.

The proportion of AS events which actually contribute to the pool of functional transcripts remains unclear as non-functional transcripts product of AS continue to be reported in multiple species ((Kalyna et al. 2011; Filichkin et al. 2010) and reviewed in (Maquat 2004; Lejeune & Maquat 2005; Keren et al. 2010; Kalsotra & Cooper 2011)).

In many cases is possible to identify if a transcript has potential to be functional by analysing its sequence. Alternative splicing can produce a frame shift in the mRNA sequence or a premature stop codon may be inserted in the sequence (Figure 1.2). A premature stop codon, product of a frame shift or if inserted as part of an alternatively spliced exon or retained intron, will potentially trigger the nonsense-mediated mRNA decay (NMD) pathway and then become target of degradation (McGlincy & Smith 2008).

It has also been suggested it is possible to predict AS functionality by observing expression patterns. A highly expressed AS event is expected to be functional. A large proportion of EST or RNA-seq data from a specific AS event will indicate there is a bias toward the conservation of the AS event, meaning that event has a high potential of having a function (Kan et al. 2002).

Other possible approach will be to study the evolutionary aspects of AS, it is expected essential AS events will be highly conserved, but this approach carries limitations, only a fraction of alternatively spliced exons are highly conserved between higher eukaryotes (Sorek et al. 2004; Modrek & Lee 2003). If a specific AS

event is responsible of a highly novel phenotype, it will be meaningless to try to trace it back in evolutionary terms. But if the AS is highly conserved and is associated with high tissue-specificity, is likely there is a function associated with that AS event (Irimia et al. 2009). Working together, tissue specificity and evolutionary conservation have been associated with frame preserving AS events, meaning there is high probability the AS event will produce a functional protein isoform ((Merkin et al. 2012) and reviewed in (Kornblihtt et al. 2013)). Nevertheless, it is important to consider some AS events control transcription regulatory factors, and accordingly, may not maintain a direct relationship with derived phenotypes, instead they may be regulating cascade reactions or the expression of other proteins in the organism.

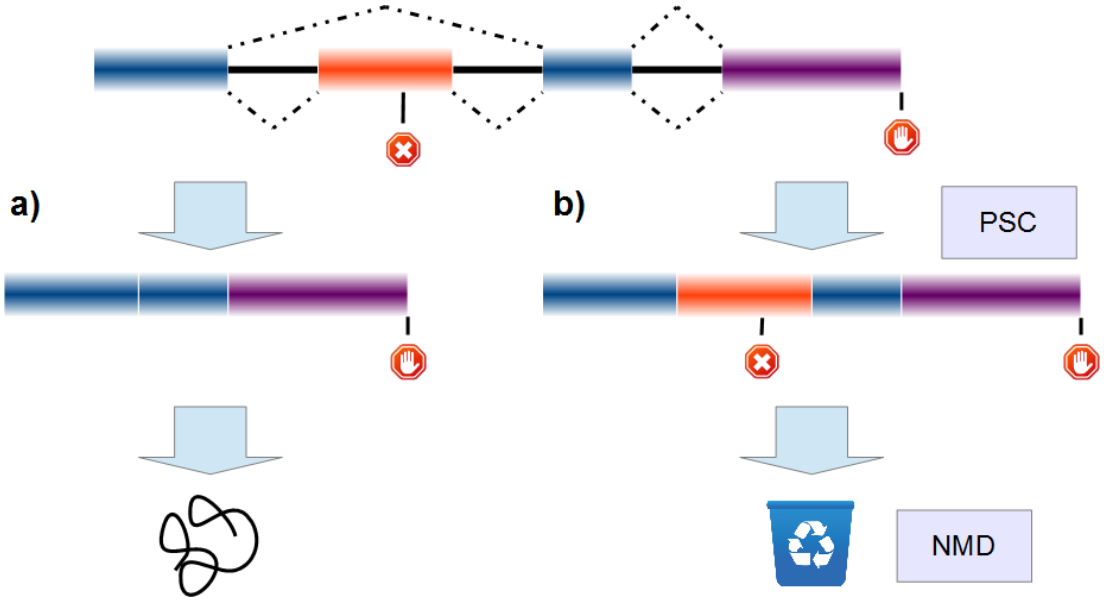


Figure 1.2. Different outcomes of alternative splicing. a) When symmetrical exons are spliced together and there is stop codon at the end of the mature mRNA, the sequence will be candidate for translation. b) A premature stop codon (PSC) is included into the mature mRNA as part of an alternatively spliced exon making it candidate for NMD.

In chapter 5, I further develop about possible strategies to assess the functional relevance of alternative splicing and its potential role in the evolution of functional genomic innovations.

1.7 Alternative splicing and disease

Due to the complexity of the AS process, several single points of failure are introduced for the steps happening before translation, changes in the *cis*-regulatory sequences of splicing can drastically change or invalidate essential gene products. Alterations in normal alternative splicing patterns are known to have a close link with disease (see (Orengo & Cooper 2007), reviewed in (Tazi et al. 2009)). Recent estimates attribute up to half of human disease mutations to changes of splice regulatory regions (Wang & Cooper 2007). There are also documented cases of disease product of overexpression of trans-acting regulation genes of AS in other mammals (Gabellini et al. 2006; Anczuków et al. 2012).

Many disease conditions associated with AS linked are small variations in transcription regulation. Deficiency of secretion of growth hormone in humans is caused by mutations in growth hormone gene. The growth hormone gene has five exons and undergoes AS normally producing two isoforms. Changes in the transcription regulation of this gene prevents the normal expression of those two isoforms (Cogan 1997; Moseley 2002) and results in under expression of growth hormone and underdevelopment of the person.

Human WT1 is a tumour suppressor gene also linked with Frasier syndrome. WT1 undergoes extensive AS but two isoforms are known to be highly conserved in vertebrates (Miles et al. 1998). Inactivation of one splice site in this genes results in the over-transcription of one of the two generally transcribed isoforms (Haber et al. 1991). This onsets a chain of developmental problems which carry urogenital anomalies and prevents WT1 from executing its tumour suppressor duties.

Humans have two nearly identical SMN producing genes, SMN1 and SMN2, both genes cooperate to maintain adequate levels of SMN protein. A single nucleotide change introduces a new regulatory element which prevents the SMN2 exons to be correctly spliced, effectively repressing the gene and producing spinal muscular atrophy (Kashima & Manley 2003).

Sub-types of Fabry disease have been found to originate as product of a single base mutation that produces either an abnormal exon skipping or the reduction of efficiency in a splice site of the α-GalA gene (Ishii et al. 2002).

The transcription factor 7-like 2 (TCF7L2) is a highly tissue specific expressed gene (Osmark et al. 2009) linked to the Wnt/β-catenin signalling pathway and closely related to transcription (Yi et al. 2005) and development (reviewed in (Peifer 2000; Moon et al. 1997)). Because of this complex relationship with several genes, TCF7L2 has been associated with many pathologies, but is especially interesting because transcription variants of this gene have been identified to strongly associate with type 2 diabetes (Grant et al. 2006). Single nucleotide polymorphisms is considered to be the a primary risk factor for type 2 diabetes (Grant et al. 2006; Helgason et al. 2007; Goodarzi & Rotter 2007; Saxena et al. 2006; Chang et al. 2007). Nonetheless, the mechanisms underlying the relationship between the single nucleotide polymorphisms and type 2 diabetes is not fully understood. A second possible explanation for the relationship between TCF7L2 and type 2 diabetes has been suggested and has alternative splicing as its foundation. A recent study (Prokunina-Olsson et al. 2009) found evidence linking type 2 diabetes and a tissue-specific AS event expressed exclusively in pancreas, pancreatic islets and colon. The expression levels for this isoform positively correlated with proinsuline production in tissue associated with non-type 2 diabetes. Even though results were not conclusive, this suggest it is this specific isoform the one regulating insulin production and changes in non-coding regions of the gene may be affecting its transcription.

Alternative splicing and cancer have been frequently associated (reviewed in (Kelemen et al. 2013; Pajares et al. 2007; Tazi et al. 2009)). The presence of aberrant alternative splicing events in cancer tissues not found in normal cells has been suggested as evidence for a functional role of these cancer specific transcripts. However, as we show in section 1 in the appendices, cancer specific transcripts tend to be specific to single cancer derived libraries with an above average presence of non-sense mutations. Thus the specific cancer promoting role of cancer specific alternatively spliced transcripts remains unclear.

On the opposite extreme, targeting regulatory elements of AS can be used as powerful tools for gene therapy. For example, both in-vivo and in-vitro models in mouse have confirmed targeting a specific intronic splicing silencer in the *SMN2* gene is an effective treatment for mild spinal muscular atrophy for this specific species and under laboratory conditions (Hua et al. 2010).

In section 1 of the appendices we investigate the characteristics of transcripts derived from cancer libraries. When comparing transcripts from cancer and normal human tissues we find cancer libraries have in general a higher proportion stop codons. This same pattern repeats when comparing alternatively spliced transcripts exclusive to cancer libraries. Also transcript from cancer libraries have fewer functional domains and have low levels of transcription. Finally when comparing alternatively spliced oncogenes with alternatively spliced tumour suppressors we find oncogenes are enriched in new functions while tumour suppressors tend to lose function.

1.8 Structure of the thesis

This thesis is organised into six chapters. In this chapter I have outlined current knowledge of alternative splicing prevalence across eukaryotic taxa and the potential functional role and contribution of genomic innovation. In chapter 2, I describe the method I have used to produce comparable prevalence of AS. In the same chapter I use the method to measure AS prevalence in protists. A eukaryote group which so far has been poorly characterised for AS. My results not only contrast previous reports of AS prevalence in protists, but also constitute the broadest analysis of AS prevalence in protists.

Another important question about AS relates to the functional roles it plays in eukaryotes. In chapter 3, I apply the same method to characterise AS in several fungi species. My results challenge previous conceptions about AS prevalence in fungi, with several fungi species showing AS prevalence comparable with lower

metazoans. I then functionally characterise alternatively spliced genes and look for possible relationships between AS and complex phenotypes in fungi species.

In chapter 4, I characterise AS in human head and body lice. These two types of lice have very similar transcriptional profiles but differentiated phenotypes. Using transcripts from two different types of high-throughput DNA sequencing systems I identified and confirmed AS events in both types of lice. I functionally characterised alternatively spliced genes. I then studied the functionality patterns looking for differences between the two types of lice. My results confirm differences between the functional characterisations of AS in head and body lice. Furthermore it was possible to associate functional differences with possible changes in phenotype.

In chapter 5, I address different questions about regulation, prevalence, functionality and evolution of AS and the role AS may be playing in functional innovation. In this last chapter I engage into the discussion about the importance of using comparable data to studying prevalence of AS. The future of AS identification using high-throughput technologies and the evolutionary road AS follow with the implications it may have to diversification and phenotypic variation.

In chapter 6, I summarise my findings and their significance in the understanding of the role of alternative splicing in eukaryotic species as well as outline a number of outstanding questions.

There are three further sections which form the appendices. These present work where I have contributed to various degrees and which further assess the role of alternative splicing in disease states in the human genome or in gene evolution in a species of brown algae Ectocarpus and polymorphic coding sequence deletions in the model plant species *Arabidopsis*.

2 Characterising alternative splicing prevalence in protist species correcting for the distorting effects of differential transcript coverage

Protists represent a large and diverse group of organisms which are ubiquitous in the global ecosystem (Corliss 2002). Protists exhibit large phenotypic diversity; they may be autotrophs or heterotrophs and constitute an important food source for larger organisms, playing a fundamental role in the food chain. Many protist species are known disease causing agents in human, domestic animals, plants and other eukaryotes. Because of the significant advantages they offer in comparison with metazoans, protists are used as model organisms to study different biological properties of eukaryotes (Montagnes et al. 2012). Nevertheless, protists are commonly overlooked in comparison with other eukaryotic organisms (Caron et al. 2009).

Alternative splicing (AS) is a regulatory post-transcriptional process which allows for the variable incorporation of protein coding regions (exon) into the mature mRNA. As a result of that variability, alternative splicing allows for multiple protein products to be encoded by a single gene (reviewed in (Breitbart et al. 1987; Graveley 2001)).

Although alternative splicing has been under intense scrutiny for the last decade, most studies have focused on human, rodent models and the fruit fly. In these species, AS has been found to be associated with a variety of fundamental physiological and developmental processes. But how AS prevalence has evolved over time along the phylogenetic tree remains poorly understood. This has been caused by several factors; first, transcript sequences are available for a relatively small number of protist species making it difficult to assess AS for a large number of protist species. In addition, few efforts have been devoted into assessing AS levels in non-metazoan or plant species. Thus, there are few examples of multiple species

characterisation of AS in protist species (McGuire et al. 2008) and those characterisations do not provide with a correction for the differential transcript coverage between genes and species (E. Kim et al. 2007).

Despite these issues, alternative splicing in protists has not been entirely overlooked, as there are a number of studies examining alternative splicing in individual species (Iriko et al. 2009; Xiong et al. 2012; Grisdale et al. 2013) or focused on specific genes (Escalante et al. 2003; Muhia et al. 2003; Singh et al. 2004). These studies have shown that alternative splicing does occur in protist species. So far only one multi-species study characterised AS in protists (McGuire et al. 2008). The study includes 9 protist species using transcript data to identify AS and found that AS was present in all protist species and per species, intron retention is the most common type of AS. Nevertheless, because levels of AS detection are highly dependent on the levels of transcript coverage (E. Kim et al. 2007), the study by McGuire et al. (2008) focused on the characterisation of AS per-species and could not draw comparative conclusions about prevalence of AS across species. Thus, questions about how prevalent is AS, how it varies across protist species and how AS prevalence in protist compares with organisms from other taxonomic groups; need to be addressed. Furthermore, generalised evidence linking AS and phenotypes in protist species remains to be found.

Here we use publicly available ESTs and the corresponding fully sequenced genomes of 18 species from various clades of protists to characterise AS prevalence. Using a transcript normalisation method we obtained comparative estimates of alternative splicing prevalence. With the AS prevalence and AS levels produced, we try to identify the functions AS is fulfilling and their relationship with protist phenotypes.

2.1 Materials and methods

2.1.1 Transcript and genome data

The genomic sequences of 18 protist species were obtained from public databases (see Supplementary table 2.1 for list of species and genomic data sources). Transcript sequences were obtained for those same 18 species from dbEST (Boguski et al. 1993) [May, 2011].

2.1.2 Alternative splicing detection

Individual AS events were identified according to the process used by Chen, Tovar-Corona, & Urrutia (2011). ESTs were aligned to their corresponding genomes using GMAP software package (Wu & Watanabe 2005). ESTs were associated to specific genes if the particular EST was uniquely aligned to a genomic region enclosed in the genomic coordinates of the gene. Only those alignments with 95% coverage and 95% identity were used for the rest of the process. ESTs were associated with the best alignment available using the ranking produced by the GMAP software. All ESTs mapping to genomic regions with no annotated protein coding genes were not considered in this study. Genes with overlapping genomic coordinates along with all associated ESTs were discarded from further analyses to avoid miss-annotation of transcription of exons of different genes as alternative splicing events. EST alignment coordinates were used to identify exon-intron boundaries allowing to obtain refined gene exon-intron templates. Genes with no ESTs associated were only considered as part of the baseline total number of genes for the prevalence studies, but were not used otherwise. The process led to the identification of non-annotated exons and elimination of orphan genes which were not included in transcripts spanning any other exon which may result from nested or overlapping genes.

Alternative splicing events were then identified by comparing the alignment coordinates of individual ESTs against the exon-intron borders as identified from all transcript alignments. ESTs were ordered from longest to shortest to identify exon-intron borders. Variations of exon-intron borders between different ESTs were

classified as potential AS events. Only those ESTs which expanded over two or more exons were considered. AS events were classified into five different types: alternative 3' splice site selection (3S), alternative 5' splice site selection (5S), alternative 3' and 5' splice sites (3S5S), exon skipping (ES) and intron retention (IR).

Each alternative splicing event's coordinates, corresponding gene and supporting EST were organised in a relational database.

2.1.3 Comparative alternative splicing indexes

Comparative alternative splicing indexes were obtained by using a transcript normalisation method (E. Kim et al. 2007). These comparative AS indexes were calculated as the average number of AS events identified in 100 replicates of 10 randomly selected ESTs from the EST pool associated with each gene. Comparable indexes were only calculated for those genes with over 10 associated ESTs. AS prevalence estimates per species were obtained by averaging comparable AS for all genes in which comparable AS indexes were calculated. Those species with comparable AS estimates were available for fewer than 50 genes were not considered (see Supplementary table 2.2 for final list of species used).

2.1.4 Alternative splicing and parasitic/free-living phenotypes

Phenotypic classifications for protist species were obtained from literature (see Supplementary table 2.3). Protist species were first divided into parasites and free living protists. Parasite protists were further divided depending on the kind of host (metazoan or plant). Third classification was based on the number of hosts required for a complete life cycle (single host or multiple hosts). For each classification, organisms were divided into two groups (one for each possible phenotype), unpaired two-sample t-tests were used to examine the comparative average number of AS events and percentage of alternatively spliced genes of the species of each group.

2.1.5 Functional characterisation of alternatively spliced genes

Gene ontology (GO) annotations were obtained for each species with 500 or more genes with comparable AS data. GO terms related to biological process

annotations were downloaded from Ensembl BioMart (Kinsella et al. 2011). Only those species with 50 or more alternatively spliced genes with GO annotations were used to functionally characterise AS. Significant over-representation of alternatively spliced genes per GO category was assessed by performing a Z–test. Expected proportions and its standard deviations for each category were obtained from 1000 Monte Carlo samples from all the genes analysed per species with Benjamini-Hochberg multiple testing corrections for the number of functional categories tested was used.

2.2 Results

2.2.1 Alternative splicing detection

In order to identify alternative splicing prevalence in protist species, over a million ESTs were aligned to their corresponding genomes. In total, 54942 genes were associated with partial ESTs in all species studied. This represents a third of the total number of non-overlapping genes in the genomes examined. These transcript-to-genome alignments were used to construct refined intron exon boundaries per gene (see methods). AS events were identified by comparing the alignment coordinates per transcript to those of the gene template. Through this method, it was possible to identify 2831 AS events in the species studied.

Notably, evidence of alternatively spliced genes was found in all of the species in the study. AS events were then classified into types (see Figure 2.1). Consistently with previous observations, we found that intron retention AS event type (IR), was the dominant type of AS in all protists species examined. Exon skipping (ES) events were found to be rare in the protist species examined. In *T. annulata*, *C. muris*, *C. parvum*, *E. tenella*, and *E. histolytica* no ES was detected.

The lack of comparable AS estimates which correct for the distorting effect of differential transcript coverage (E. Kim et al. 2007) has hampered the study of alternative splicing prevalence among protist species. In order to address this issue,

comparable AS indexes were obtained using a transcript normalisation method where alternative splicing was calculated per gene from samples of a set number of transcripts per gene (see methods). Comparing the pre- and post-normalisation data for protists (Figure 2.2) confirms our method corrects the bias for higher AS level detected in genes with ample transcript coverage.

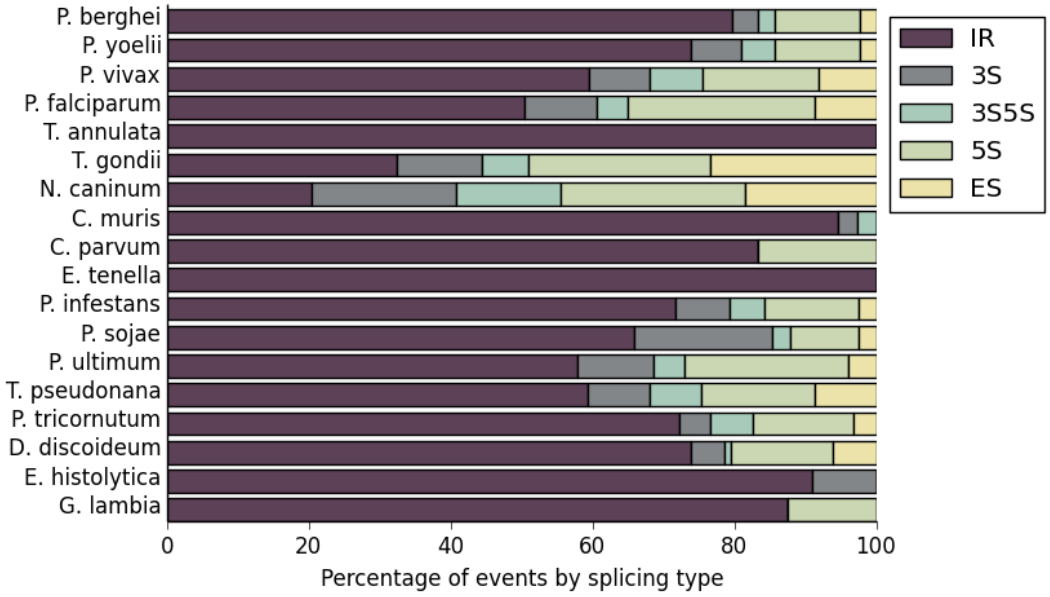


Figure 2.1. Proportion of AS by event type in protists. AS events were classified in 5 different types, intron retention (IR), alternative 3' splice site selection (3S), alternative 3' and 5' splice site selection (3S5S), alternative 5' splice site selection (5S) and exon skipping (ES). Figure uses non-comparative AS data so only species specific observations apply. In general IR is the dominant type of AS event in protists. But individual species show high variability, in some cases ES represents close to 20% of the total events detected. Our data shows high prevalence of 5S AS events in *P. falciparum* among several other species, this confirms previous observations for *P. falciparum* (Iriko et al. 2009; McGuire et al. 2008) but extends the pattern to other distant related protists.

Comparable data was obtained for 6308 genes across all species. Of those, 749 genes had evidence of AS (see Figure 2.3 and Supplementary table 2.2 for detailed information about each species). Using these estimates AS prevalence and average number of AS events per gene were calculated for each species. All those species where comparable indexes were obtained for fewer than 50 genes were removed from all comparable AS levels analysis. On average, about 11.34% of genes were found to be alternatively spliced. However, there were marked variations in

prevalence among species with the highest prevalence of AS detected in N. caninum and P. vivax (30% and 31%).

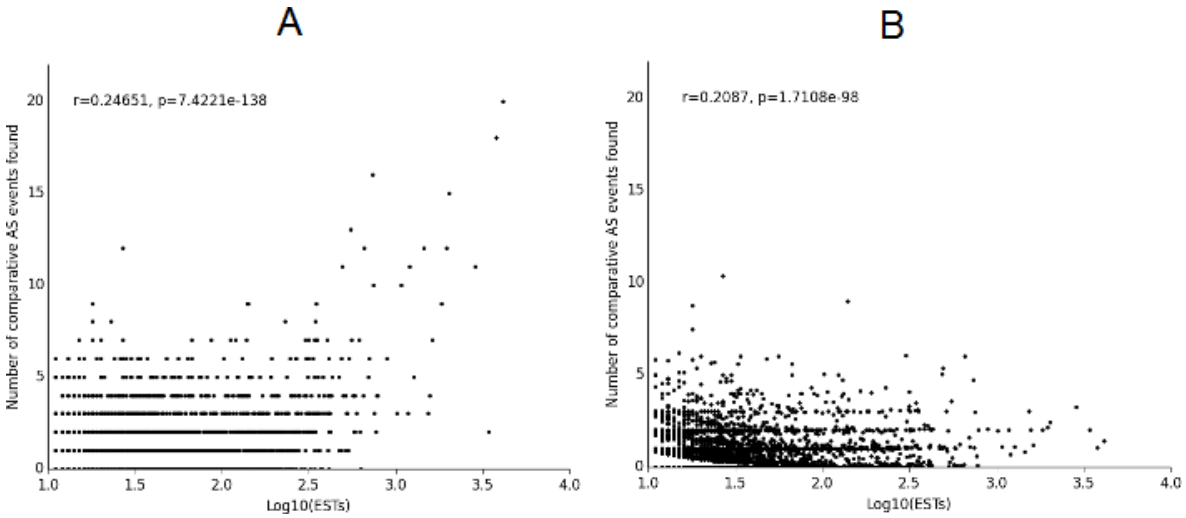


Figure 2.2. [A] Because higher number of EST allow for the detection of higher levels of splicing there is known bias toward the detection of higher levels of AS in species with lager databases of transcript sequences (E. Kim et al. 2007; Kan et al. 2002). Figure shows the whole set of genes with more than 10 ESTs available and the corresponding number of AS detected in each gene for all protist species. Is possible to identify that there is a strong association between the levels of AS detected and the number of ESTs available. [B] Effect of using an EST coverage normalisation method effectively counteracts the effect of the EST coverage on the AS levels detected. This correction allows for the effective comparison of AS levels between species when is not possible to warranty same transcript coverage for different species (E. Kim et al. 2007).

Notably, prevalence detected for these two species surpasses any previously reported prevalence of AS in protists (Xiong et al. 2012) and is comparable with higher eukaryotes (Figure 2.5). Comparing average AS event number per species, we find that protist have high variability. In average protists genes have 0.27 AS events per gene (Figure 2.3 and supplementary table 2.2) and there is a positive correlation between AS prevalence and AS levels (Figure 2.4).

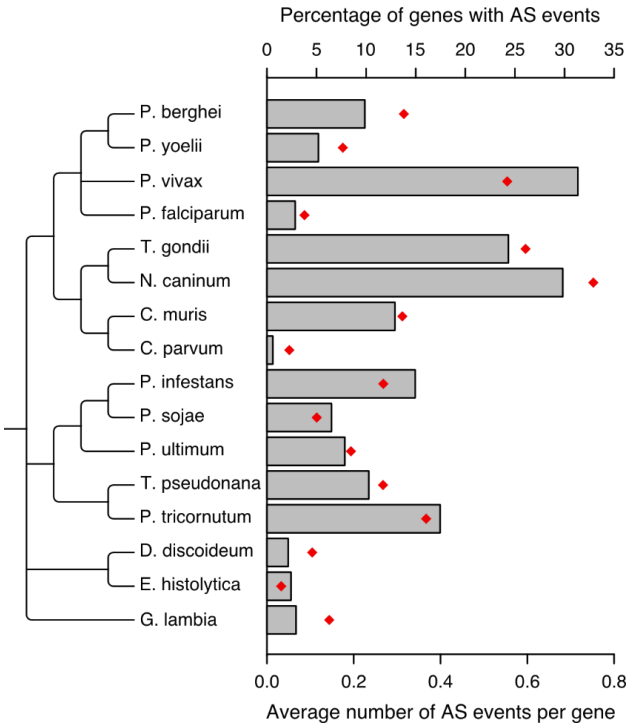


Figure 2.3. Showing alternative splicing prevalence and levels of AS detected from EST sequences. Bars shows the percentage of genes with alternatively spliced transcripts using comparative data. Red dots represent the average number of AS events. Only those species with more than 50 genes with comparable data are shown (16 species).

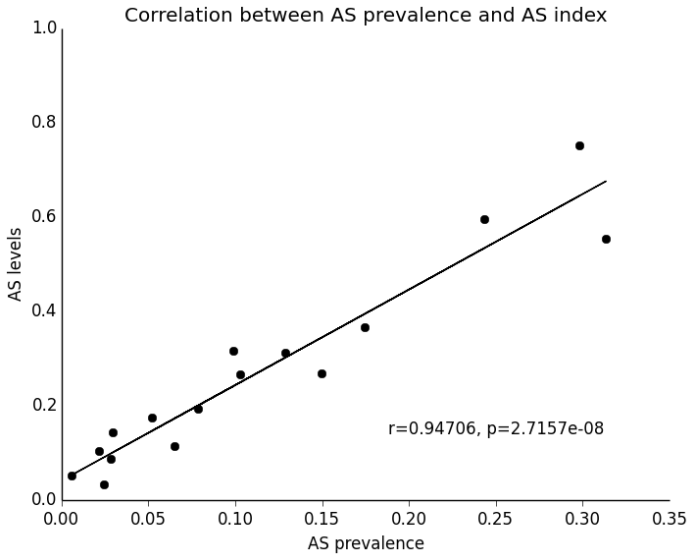


Figure 2.4. Showing the positive correlation (Spearman rank correlation) for AS prevalence and AS levels on the species studied. Only those species with more than 50 genes with comparable data are shown (16 species).

In order to assess whether observed variations in AS events among protist species followed a phylogenetic pattern with more closely related species having more similar AS prevalence, we used a phylogenetic generalized least squares (PGLS) test to assess phylogenetic signal on the prevalence of AS across the analysed fungal species. We uncovered no evidence for a phylogenetic association in AS prevalence or average AS levels across species ($p > 0.05$).

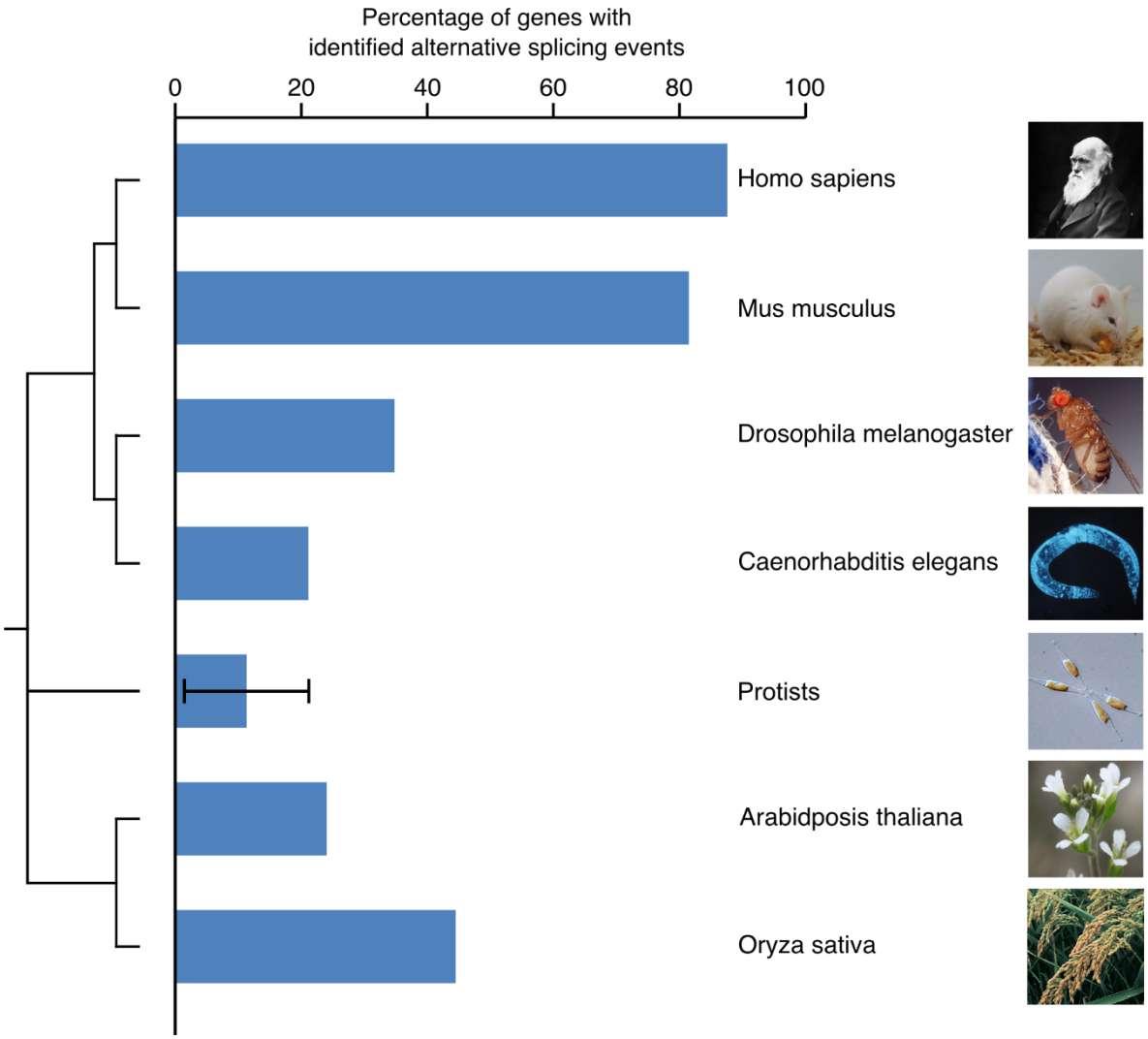


Figure 2.5. Comparative AS prevalence in eukaryote species (cladogram from (Sayers et al. 2009)). Error bar represents fluctuation (standard deviation) of AS prevalence for protist species (16 species in sample). Figure shows individual protist species have similar comparative prevalence of AS as *C. elegans*.

2.2.2 Alternative splicing and parasitic/free-living phenotypes

As alternative splicing has been proposed to have an impact in the proteome size and potential functional diversification at the genomic level, we examined whether alternative splicing prevalence among the protist species examined could be linked with various complex phenotypes (see methods). We examined whether a species was a parasite or a free living cell. We also assessed differences among parasites according to the nature of their host, plant or metazoan. Finally we also looked at differences in AS prevalence and levels comparing the number of steps necessary to complete the life cycle of the species. We found however, no significant differences between AS levels in parasite opposed to free-living protists ($p = 0.764$ for average number of AS events and $p = 0.772$ percentage of alternatively spliced genes). For parasite protist species, no significant difference was found for AS parameters in plant parasites compared to metazoan parasites ($p = 0.606$ for average number of AS events and $p = 0.258$ percent of alternatively spliced genes) nor between protists with direct and indirect life cycles ($p = 0.114$ for average number of AS events and $p = 0.067$ for percentage of alternatively spliced genes). These results show that we failed to find any evidence linking various complex phenotypes and AS prevalence.

2.2.3 Functional characterisation of alternatively spliced genes

To better understand the potential functional role of alternative splicing of alternative splicing in cellular processes, we examined the functional gene ontology (GO) term associations of alternatively spliced genes and the wider gene pool. Due to sample size constraints just two species were available to be functionally characterised. Only *P. tricomutum* and *P. infestans* complied with both the minimum number of genes with comparative AS data and the minimum number of genes with functional GO term annotations. Of the two species only *P. infestans* was enriched on translation ($p < 0.00001$; see Figure 2.6). No functional category was significantly under-represented.

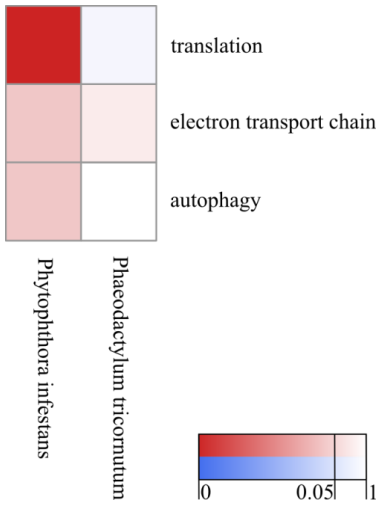


Figure 2.6. Showing GO categories which appear related to AS in a higher proportion as compared to a random sample of genes. Alternative spliced genes in *P. infestans* are specifically enriched in the translation GO category. The scale shows how statistically significant is the result, both for enrichments (red) and impoverishments (blue). Intensity of colour represent degree if statistical significance. Although three GO categories were found to be enriched for *P. infestans*, only translation was statistically significant. Enrichment was tested also for *P. tricornutum* but no evidence for enrichment was statistically significant.

2.3 Discussion

Protists represent a large and diverse set of organisms and have wide variations at the genomic level with some protists having extraordinarily big genomes (Hackett et al. 2004; Bachvaroff & Place 2008) and a propensity to have gene transfer (Bachvaroff & Place 2008; Lowe et al. 2011). Alternative splicing has been widely studied in human and other metazoan model organisms. Alternative splicing has been reported for a number of protist genomes but it is difficult to infer AS prevalence directly from genomic data. Previous studies in protozoa generally rely on transcription data to identify or characterise AS limiting the number of species analysed (Xiong et al. 2012; Lowe et al. 2011; Muhia et al. 2003; Coyne et al. 2008).

Here we have presented the first comparable study of AS prevalence in protists species. AS detection methods which rely in transcription data to identify AS events are biased toward species with high transcript coverage (Kim et al. 2004; Brett et al. 2002). Using available transcript databases and a sampling method which corrects for transcript coverage bias, we have produced comparable AS prevalence indexes for 16 protist species with fully sequenced genomes. Our results show previous reports underestimate AS prevalence in protists species. The highest prevalence so far reported for a protist species (*T. thermophila* 5.2% (Xiong et al. 2012)) greatly contrasts with the AS prevalence corrected for transcript coverage. We found that if we take into consideration all protists species, more than 11 % of protists genes undergo AS. Moreover our results show at least two species have close to 30% of AS prevalence. Such prevalence is just as high as the prevalence calculated for *C. elegans* using the same method to correct for transcript coverage. Our results also show high variability in AS prevalence although is important to take into consideration the study includes species coming from a diverse selection of protists groups expanding some 1300 million years of evolution (Hedges et al. 2004).

A high prevalence in protists is relevant because in multicellular organisms, where different cells types require to express different transcripts AS may be playing an important role for tissue specific gene expression (reviewed in (Matlin et al. 2005)); it has been proposed (Xiong et al. 2012) that in protists, alternative spliced genes may play a similar role but as a subcellular-structure-specific promoter of gene function diversification. AS has been shown to be specific to stages of the life cycle in *P. falciparum* (Muhia et al. 2003) and *T. thermophile* (Xiong et al. 2012) consistent with a functional role of alternative splicing transcripts. Our results show that alternatively spliced genes are disproportionally associated with translation in one of the two species analysed, however we found no evidence of an overall association of alternative splicing prevalence among protist species and a parasitic lifestyle or the complexity of its life cycle.

An over-representation of the transcription GO category in genes associated with AS suggests AS may be playing a functional role in protists. Previously AS events conserved between human and mouse have been found to be associated with

the transcription regulation category (Yeo et al. 2005). One possible explanation will be for AS to be playing a functional role as part of gene interactions, making it more difficult to pinpoint specific enriched functions for alternatively spliced genes which result in phenotype variations since they may be working as part of a cascade effect.

2.4 Conclusions

This study implements the fundaments of a transcript normalisation protocol to study AS. Here we describe the steps taken to produce comparative AS prevalence indexes using transcript and genome data from eukaryote species. The process takes into consideration the transcript coverage bias which affects AS characterisation when using transcript databases.

Using the process developed, we also present a characterisation of AS in 18 protist species. To the best of our knowledge this is the most comprehensive study of AS prevalence in protists, including 16 species with data for comparable AS prevalence. Per species analysis shows intron retention is the most common AS type in protists. Our results show 11 percent of protist genes are alternatively spliced. But prevalence is variable between species. This results add to the discussion about AS prevalence as is the highest prevalence so far reported for protists. This study also shows even with the data limitations is possible to identify statistically significant relationships between genes undergoing AS and functions related to DNA translation in at least one of the species studied.

2.5 Supplementary tables

Supplementary table 2.1: Protist species included in the study.

Number of transcripts used per species to detect AS events and the genomic data sources.

Species	Number of transcripts	Database and genome version
C. muris	27498	EupathDB (Aurrecoechea et al. 2010)
C. parvum	60450	EupathDB (Aurrecoechea et al. 2010)
D. discoideum	155032	Ensembl 62 (Kinsella et al. 2011), Dictybase.01.1a
E. tenella	35009	Sanger (Ling et al. 2007)
E. histolytica	20812	EupathDB (Aurrecoechea et al. 2010)
G. lamblia	20238	EupathDB (Aurrecoechea et al. 2010)
N. caninum	25072	EupathDB (Aurrecoechea et al. 2010)
P. tricornutum	133887	JGI (Grigoriev et al. 2012), Phatr_JGIv2
P. infestans	111106	Ensembl 62 (Kinsella et al. 2011), Phyinf1_1.1
P. sojae	28467	Ensembl 62 (Kinsella et al. 2011), Physo1_1.1
P. berghei	23782	EupathDB (Aurrecoechea et al. 2010)
P. falciparum	53293	EupathDB (Aurrecoechea et al. 2010)
P. vivax	31855	EupathDB (Aurrecoechea et al. 2010)
P. yoelii	16254	EupathDB (Aurrecoechea et al. 2010)
P. ultimum	100391	Ensembl 62 (Kinsella et al. 2011), Pu1.1
T. pseudonana	61913	Ensembl 62 (Kinsella et al. 2011), Thaps3.1
T. annulata	17031	Sanger (Pain et al. 2005), TANN_092304
T. gondii	136229	EupathDB (Aurrecoechea et al. 2010)
A. thaliana	1529700	PlantGDB (Duvick et al. 2008)
C. elegans	396687	Ensembl (Kinsella et al. 2011)
D. melanogaster	821005	Ensembl (Kinsella et al. 2011)
H. sapiens	8315122	Ensembl (Kinsella et al. 2011)
M. musculus	4853546	Ensembl (Kinsella et al. 2011)
O. sativa	1252989	PlantGDB (Duvick et al. 2008)

Supplementary table 2.2: Prevalence of AS for protists using comparable data

Detail of AS prevalence per species. Second column shows the number of genes with enough data for comparative AS prevalence detection.

Species	Total genes with more than 10 EST	Number of genes with evidence of AS	Percent of genes with evidence of AS	Comparative AS events	Average number of AS events per gene (considering all genes with comparative data)
C. muris	62	8	13%	19.35	0.3120968
C. parvum	169	1	1%	8.73	0.0516568
D. discoideum	790	17	2%	82.46	0.1043797
E. histolytica	206	5	2%	6.78	0.0329126
G. lamblia	136	4	3%	19.53	0.1436029
N. caninum	57	17	30%	42.87	0.7521053
P. tricornutum	504	88	17%	184.94	0.3669444
P. infestans	669	100	15%	179.45	0.2682362
P. sojae	261	17	7%	29.97	0.1148276
P. berghei	152	15	10%	47.96	0.3155263
P. falciparum	807	23	3%	69.85	0.0865551
P. vivax	370	116	31%	204.88	0.5537297
P. yoelii	154	8	5%	26.95	0.175
P. ultimum	751	59	8%	145.44	0.1936618
T. pseudonana	185	19	10%	49.5	0.2675676
T. gondii	1035	252	24%	616.83	0.595971

Supplementary table 2.3: Parasitic and free-living protist phenotypes

Classification for parasitic and free-living protists showing also the number of hosts required for a complete life cycle and the type of host associated with the parasite.

Species	Parasite/free-living	Host type	Number of hosts
C. muris	Parasite (Katsumata et al. 2000)	Metazoan (Palmer et al. 2003; Ramirez et al. 2004)	Single (Ramirez et al. 2004)
C. parvum	Parasite (DuPont et al. 1995)	Metazoan (Ramirez et al. 2004)	Single (Ramirez et al. 2004)
D. discoideum	Free-living (Bozzaro & Eichinger 2011)	NA	NA
E. histolytica	Parasite (Boettner et al. 2008; Pham Duc et al. 2011)	Metazoa (Haque et al. 2003)	Single (Haque et al. 2003)
G. lamblia	Parasite (Cacciò & Ryan 2008; Jerlström-Hultqvist et al. n.d.)	Metazoa (Cacciò & Ryan 2008)	Single (Cacciò & Ryan 2008)
N. caninum	Parasite (McAllister et al. 1998)	Metazoa (McAllister et al. 1998)	Multiple (McAllister et al. 1998)
P. tricornutum	Free-living (Scala et al. 2002)	NA	NA
P. infestans	Parasite (Shattock 2002)	Plant (Shattock 2002)	Single (Shattock 2002)
P. sojae	Parasite (Tyler 2007)	Plant (Waugh 2000)	Single (Waugh 2000)
P. berghei	Parasite (Miller et al. 2002; Michel et al. 2005)	Metazoa (Sturm et al. 2006)	Multiple (Sturm et al. 2006)
P. falciparum	Parasite (Miller et al. 2002; Florens et al. 2002)	Metazoa (Bousema & Drakeley 2011)	Multiple (Bousema & Drakeley 2011)
P. vivax	Parasite (Miller et al. 2002; Carlton et al. 2008)	Metazoa (Bousema & Drakeley 2011)	Multiple (Bousema & Drakeley 2011)
P. yoelii	Parasite (Carlton et al. 2002)	Metazoa (Baer et al. 2007)	Multiple (Baer et al. 2007)
P. ultimum	Parasite (Campion et al. 1997)	Plants (Lumsden 1976)	Single (Lumsden 1976)
T. pseudonana	Free-living (Armbrust et al. 2004)	NA	NA
T. gondii	Parasite (Taylor et al. 2006)	Metazoa (Dubey 2004)	Multiple (Dubey 2004)
T. annulata	Not enough AS data for analyses		
E. tenella	Not enough AS data for analyses		

3 Alternative splicing prevalence in fungal species

3.1 Introduction

Alternative splicing (AS) is a post-transcriptional process by which multiple functional mRNAs are produced from a single gene by selectively cutting out segments from pre-RNA molecules (Berget et al. 1977; Chow et al. 1977). This process is now considered ubiquitous among eukaryotic taxa and has been under increasing scrutiny as it has the potential to expand the proteome beyond the limitations of gene number and has been proposed as a major contributor to organism complexity (Nilsen & Graveley 2010). While intensive characterisation in human has revealed that up to 95% of genes are alternatively spliced (Pan et al. 2008), prevalence in other taxa, has been less well characterised. Alternative splicing has been particularly understudied in fungi species, partly stemming from the fact that AS is very rare in the prime fungi model species *Saccharomyces cerevisiae* with less than 5% of its genes being alternative spliced (McGuire et al. 2008; Hirschman et al. 2006). Moreover, in *S. cerevisiae* the lack of important molecular components of the splicing machinery (Kupfer et al. 2004) led some to believe that AS was rare among all fungi species.

Over the last decade, AS has been studied in an increasing number of fungi species including: *Cryptococcus neoformans* (Loftus et al. 2005), *Magnaporthe grisea* (Brown et al. 2008), *Fusarium graminearum* (Zhao et al. 2013) and *Aspergillus flavus* (Chang & Muddiman 2011).These studies have shown that AS is a common feature in fungal genomes. Studies of AS in fungi species, however, tend to be species specific (Chang & Muddiman 2011; Loftus et al. 2005; Hirschman et al. 2006; E. Kim et al. 2007; Zhao et al. 2013), focus on a limited number of genes (Baba et al. 2005; Brown et al. 2008; Marshall et al. 2013), or were not formulated to provide comparable estimates for AS prevalence (McGuire et al. 2008). Estimates of AS, are not comparable between species because of the known distorting effect of

differences in transcript coverage among genes and species on AS event detection (E. Kim et al. 2007; Kan et al. 2002; Chen et al. 2012). Thus, estimates of AS prevalence can vary between studies as more transcript sequences become available. For example, the prevalence of AS in *A. flavus* was originally estimated to affect about 1.6% of its genes (Chang et al. 2010) but this was revised to 15.4% of genes after using a different transcript set than that used for the previous estimate (Chang & Muddiman 2011). Thus, how AS prevalence compares among different fungi species and its potential association with complex phenotypes remains unclear.

Here we assess AS prevalence and average number of AS events per gene in 23 fully sequenced fungal species for which partial transcripts ESTs data are available. By applying a transcript number normalisation method (E. Kim et al. 2007) in the 19 species with the highest EST coverage per gene, we obtained comparable estimates of AS levels (number of AS events per gene) and prevalence (proportion of genes with AS events). We examined AS patterns among fungal phylogenetic groups and assessed the relationship between AS levels and a number of phenotypes associated to organismal complexity.

3.2 Methods

3.2.1 Genome and transcript sequences

Genome sequence and gene annotations where downloaded for 23 species, 18 belonging to *Ascomycota*, 3 to *Basidiomycota* and 2 to the *Zygomycota* phylum (see Supplementary table 3.1 for full list of species and data sources). Overall, 245591 genes where identified from the original gene annotations in all species. Genes with overlapping genomic coordinates were discarded from the analysis. A total of 1154107 EST sequences corresponding to any of the 23 species studied were obtained from dbEST (Boguski et al. 1993) [downloaded May 2011].

3.2.2 Alternative splicing event detection

Alternative splicing events were identified following the method used by Chen et al. (2011). In brief, EST sequences were aligned to their respective species genome using Gmap software (Wu & Watanabe 2005). Particular ESTs were assigned to a gene if its aligning coordinates fell within the outer boundaries of a gene's coding sequence in their respective species. In total 77223 genes (31.44% of the genes annotated) for all the fungal species studied had at least one associated EST. The alignment coordinates of all ESTs to their respective gene were used to refine annotations of exon intron boundaries. This allowed to discard a small number of orphan exons or annotate previously un-annotated exons. Alignment coordinates for each EST were then compared to the exon-intron boundary coordinates to identify AS events.

Comparative AS indexes were obtained by using a transcript number normalisation procedure based on previous work by E. Kim et al. (2007), calculating AS events per gene as the average number of AS events identified in 100 samples of 10 randomly selected ESTs. A total of 19 species for which comparable AS indexes could be calculated (those species with more than 50 genes with comparable data) where used for the comparative AS analyses (see Supplementary table 3.3 for final list of species included in comparative analyses).

To determine whether the distribution of the prevalence of genes with AS events across the species of fungi depends significantly on phylogenetic relatedness, we determined the magnitude of Pagel's λ (Pagel 1999), a measure of phylogenetic signal, using the R package "caper" (Orme et al. 2012). Taxonomic tree of the species was obtained from NCBI's taxonomy database (Sayers et al. 2009). Divergence time between species was obtained from the TimeTree database (Hedges et al. 2006). Where there was no estimate of divergence time between species the terminal branches where grouped into clades and the average prevalence of genes with ASEs in those clades where calculated and used for the model.

3.2.3 Functional characterisation of genes

Gene ontology annotations were obtained for *N. crassa* genes from Ensembl BioMart (Kinsella et al. 2011). Significant over-representation of alternatively spliced genes per category was assessed by performing a Z–test. Expected proportions and its standard deviations for each category were obtained from 1000 Monte Carlo samples from all the genes analysed per species. Benjamini-Hochberg multiple testing corrections against the number of categories tested in each analysis was done.

3.2.4 Multicellularity

Single celled or multicellular status for each species was assessed from literature (supplementary table 3.4). Please note that some species, even if classified as single celled, have been reported to form multicellular structures under certain conditions. In the case of *U. maydis* it was possible to classify ESTs originated from filamentous (LIBEST_014403, LIBEST_014404, LIBEST_014457) or from haploid cell samples (LIBEST_020632, LIBEST_020633, LIBEST_020634).

3.3 Results and discussion

3.3.1 Alternative splicing events identified in all fungi species analysed

To characterise AS prevalence among fungal species, AS events were identified in 23 fully sequenced genomes of fungi species using over one million available partial transcripts or ESTs (see methods, Supplementary table 3.1). We identified a total of 5614 AS events across all species. When examining individual AS event types, intron retention (IR) was found to be the most prevalent type of AS event in all species examined (Figure 3.1, Supplementary table 3.2). This is consistent with previous reports where IR was the most prevalent type of AS event type in fungi and plant genomes (McGuire et al. 2008; Kempken 2013; Kim et al. 2008a; E. Kim et al. 2007; Sugnet et al. 2004; Ner-Gaon et al. 2004) and in contrast

to metazoan species, where exon skipping (ES) has been reported to be the most common type of AS event (Kim et al. 2008a; McGuire et al. 2008; Sugnet et al. 2004).

When comparing the relative proportions of AS events across species, we observed marked variations. While in some species, such as *Trichoderma atroviride* and *Coccidioides posadasii*, IR represents more than 70% of the detected AS events, in others, non-IR AS represent more than a third of the events detected (Figure 3.1). For example, in *Melampsora laricis-populina* IR represents less than half of the total events. However, we observed some disparities among seven species in common with a previous report (McGuire et al. 2008). Notably we found that *S. pombe* is not exclusively associated with IR, having just under 70% of AS using IR.

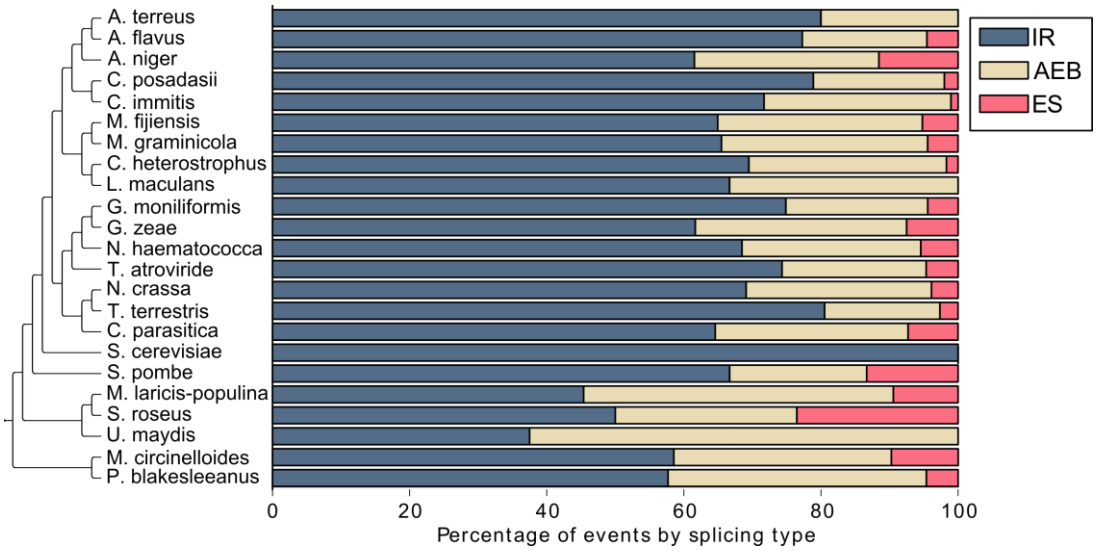


Figure 3.1 showing distribution of AS event types as a percentage of the whole AS event pool for each species. AS events were classified in 3 groups, intron retention (IR), exon skipping (ES) and alternative exon boundary (AEB). Figure uses non-comparative AS data, only individual species observations apply. IR is the dominant type of AS in fungi, but only in one species AS is exclusively associated with IR. For the rest of the species non-IR AS events represent more than 20% of the events detected in each case.

These analyses show that AS events were recovered in all species analysed, suggesting this post-transcriptional process might be a universal or nearly universal feature of fungal genome. Although, as previously mentioned, IR seems to be the

most common AS type in all species analysed, strong variations were observed in the relative proportions of each AS type across species. Similar strong variations were previously observed by McGuire et al. (2008) on a different set of fungi species, but in contrast, the present study shows higher proportions of ES in several species and also associates different types of AS event for species previously exclusively associated with IR (*S. pombe*).

3.3.2 Variable prevalence of alternative splicing among fungi species

Alternative splicing events have the potential to expand the protein pool of a species (Nilsen & Graveley 2010). However, the prevalence of AS across fungal species remains poorly understood. This is partly the result of the distorting effects of differential transcript coverage where higher transcript number results in higher AS event detection rates (E. Kim et al. 2007; Brett et al. 2002; Nilsen & Graveley 2010). To assess the prevalence (proportion of alternatively spliced genes) and level (number of AS events per gene) in fungi species, we used a transcript normalisation method (E. Kim et al. 2007). Comparative AS indexes were obtained for the 19 species where EST coverage allowed comparable AS to be calculated for at least 50 genes (see methods). We found that on average, 18.75% of fungal genes are alternatively spliced. However, when examining individual fungi species, we observed a highly variable prevalence of alternative splicing across different species ranging from less than 10% to over 30% (Figure 3.2; Supplementary table 3.3). A similarly high variability was observed in the number of AS events per gene with an average of 0.25 per gene (Figure 3.2, Supplementary table 3.3). Among alternatively spliced genes only, we observed an average of 1.52 AS events per gene. We found no significant differences in AS prevalence or average number of AS events per gene when comparing *Basidiomycota* and *Ascomycota* ($p = 0.840$ for number of genes with AS detected and $p = 0.899$ for average number of AS per gene).

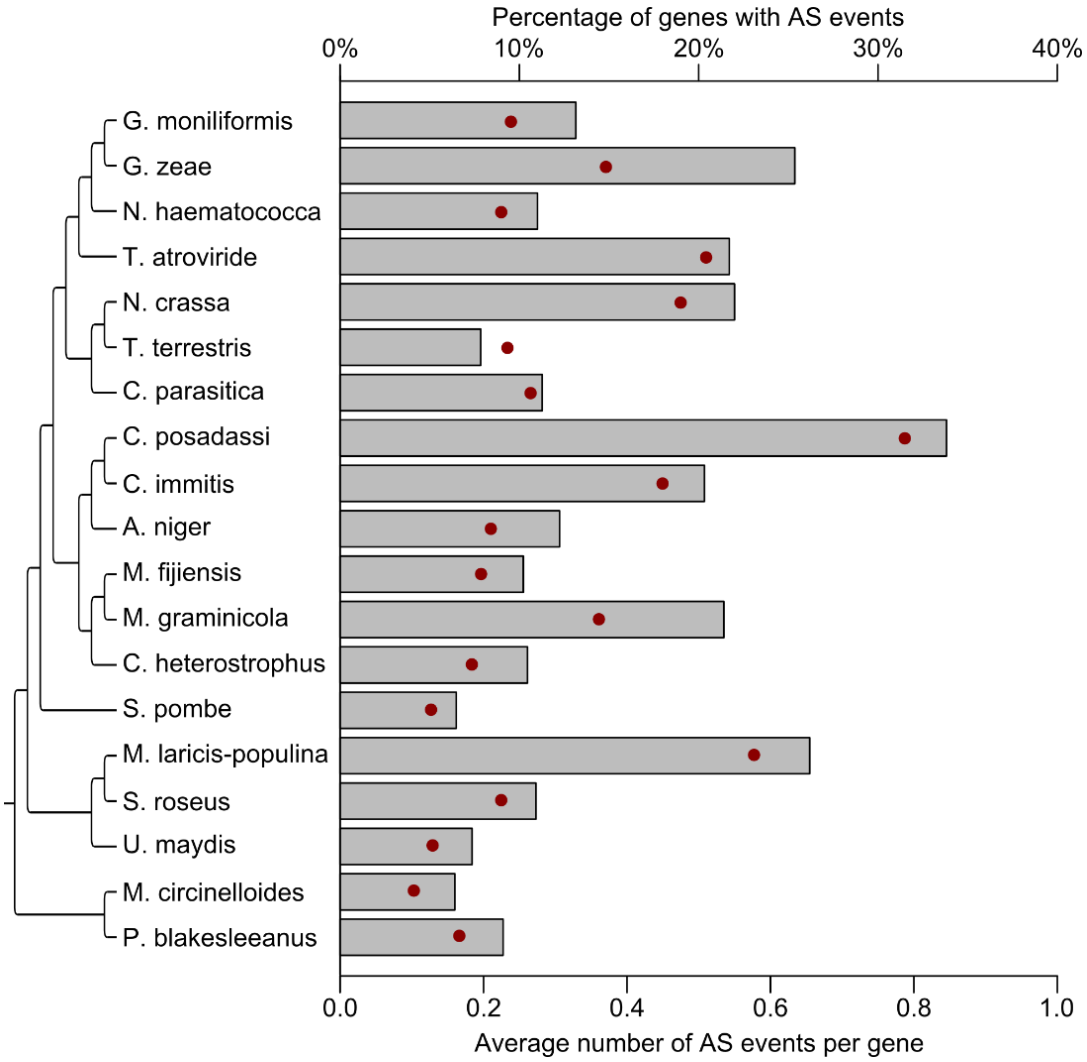


Figure 3.2. Alternative splicing prevalence and levels in fungal genomes. Bars shows the percentage of genes classed as alternatively spliced using comparative indexes which correct for the distorting effect of differences in transcript coverage between species. Red dots represent the average number of AS events for genes with sufficient EST coverage. Only 19 species with comparative indexes available for at least 50 genes are shown. Fungal species are ordered according to phylogenetic relationships.

By applying the same methods to model non-fungi species we were able to compare the prevalence of AS in fungi species from prevalence of AS found in model non-fungal species. We found that while, on average, AS prevalence among fungi species is lower compared to model species like the plant *Arabidopsis*, the invertebrates *C. elegans* and the fruit fly as well as the vertebrates human and mouse, some fungi species have similar or even higher AS prevalence to those found in invertebrates (Figure 3.3).

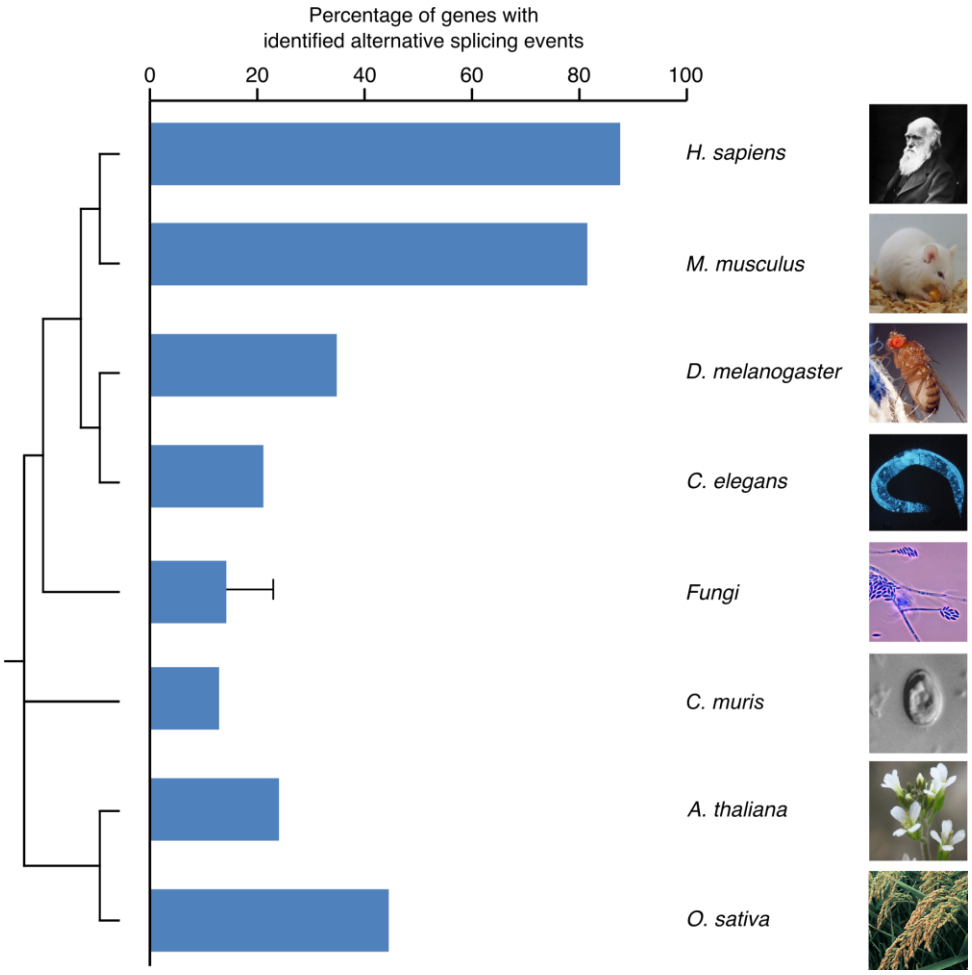


Figure 3.3. Prevalence of AS in fungi and representative model eukaryotic species (cladogram (Sayers et al. 2009)). Error bar in fungi represents the fluctuation (standard deviation) of AS prevalence detected in 19 fungi species in the sample. Figure shows there are fungi species with AS levels comparable with non-fungi species (*C. elegans* and *C. muris*) when using comparable AS data.

3.3.3 Functional characterisation of alternatively spliced genes

In order to assess the potential functional role of AS in fungal genomes we analysed functional gene ontology annotations, GO terms, in *N. crassa*, the species with the highest number of genes with both comparable AS index and GO term annotations. Using a randomisation method, we calculated over and under representation of alternatively spliced genes in individual GO categories compared to the wider gene pool. We found significant enrichments among broad functional categories including translation, biosynthetic process and protein metabolic process (Figure 3.4).

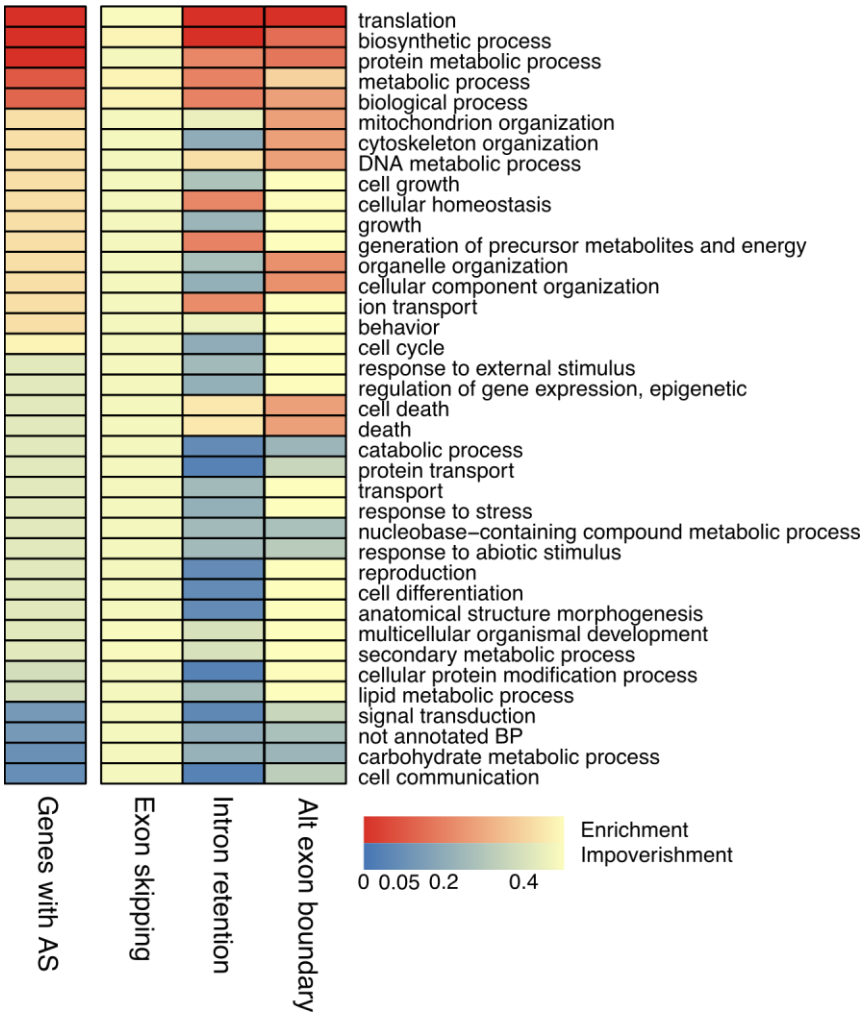


Figure 3.4. Heat map diagram depicting the over representation of biological process GO terms categories (gene ontology terms) among alternatively spliced genes in *N. crassa*. Enrichment of a GO term means it appears to be related to a higher than expected number of genes with AS in contrast with a background of all genes with comparative data. Enrichment was tested for *N. crassa*. Significant enrichments are observed in translation, biosynthetic process and protein metabolic process categories.

Because it has been suggested that different alternatively splicing event types are differentially associated with gene functions, we re-examined GO term representation for each AS type. Using comparable estimates for each AS type, once again we found statistically significant enrichment/impoverishment for certain GO terms (Figure 3.4). As expected, given that it constitutes the most common AS event type, IR showed a similar enrichment pattern as the one found in the analyses of all AS types pooled with significant enrichments in translation and biosynthetic process.

No GO terms were found to be enriched among genes with exon skipping possibly because of the small number of genes associated with this AS type.

These results show that alternative splicing is not randomly distributed among genes, but is instead more closely associated with certain GO terms in the *N. crassa* genome, particularly among genes associated with translation.

3.3.4 Alternative splicing levels are associated with multicellularity

Alternative splicing has been proposed to play an important role in functional genomic innovation (reviewed in (Chen et al. 2012)) and has been suggested as a major contributor of organism complexity by boosting proteome sizes (Koralewski & Krutovsky 2011; Graveley 2001; Nilsen & Graveley 2010; Ast 2004; Schad et al. 2011). Thus, we examined the possible link between AS and fungal multicellularity (see Supplementary table 3.4 for species classification).

Having comparable AS estimates allowed us to compare AS prevalence and levels in single celled fungi species against multicellular species. We found a significant difference in both AS prevalence and levels with multicellular species being associated with higher AS (t-test: AS prevalence: $p = 0.006$ and AS levels $p = 0.043$; Figure 3.5, Supplementary table 3.5). This result is consistent with the diversification of the transcript pool in those species forming multicellular structures in the fungal taxa. However, it is possible that these differences could result from an association between both AS prevalence and multicellularity along the fungi phylogenetic tree. In order to rule out the possibility of the observed pattern being an artefact caused from phylogenetic relatedness we applied a PGLS test to assess the strength of the phylogenetic signal in alternative splicing prevalence and levels. We observed no phylogenetic signal on the prevalence of AS across the analysed fungal species (Pagel's $\lambda = 0$, p-value $_{\lambda=0} = 1$, p-value $_{\lambda=1} = 0.027442$) or when examining AS levels per gene across taxa (Pagel's $\lambda = 0$, p-value $_{\lambda=0} = 1$, p-value $_{\lambda=1} = 0.018899$). Thus, the observed variations in AS prevalence and levels among are not explained by the degree of phylogenetic relatedness which is perhaps not surprising given that distances between species analysed can span over 400 million years

(Taylor & Berbee 2006) and alternative splicing has been observed to be a relatively fast evolving trait in other eukaryotes (Ermakova et al. 2006).

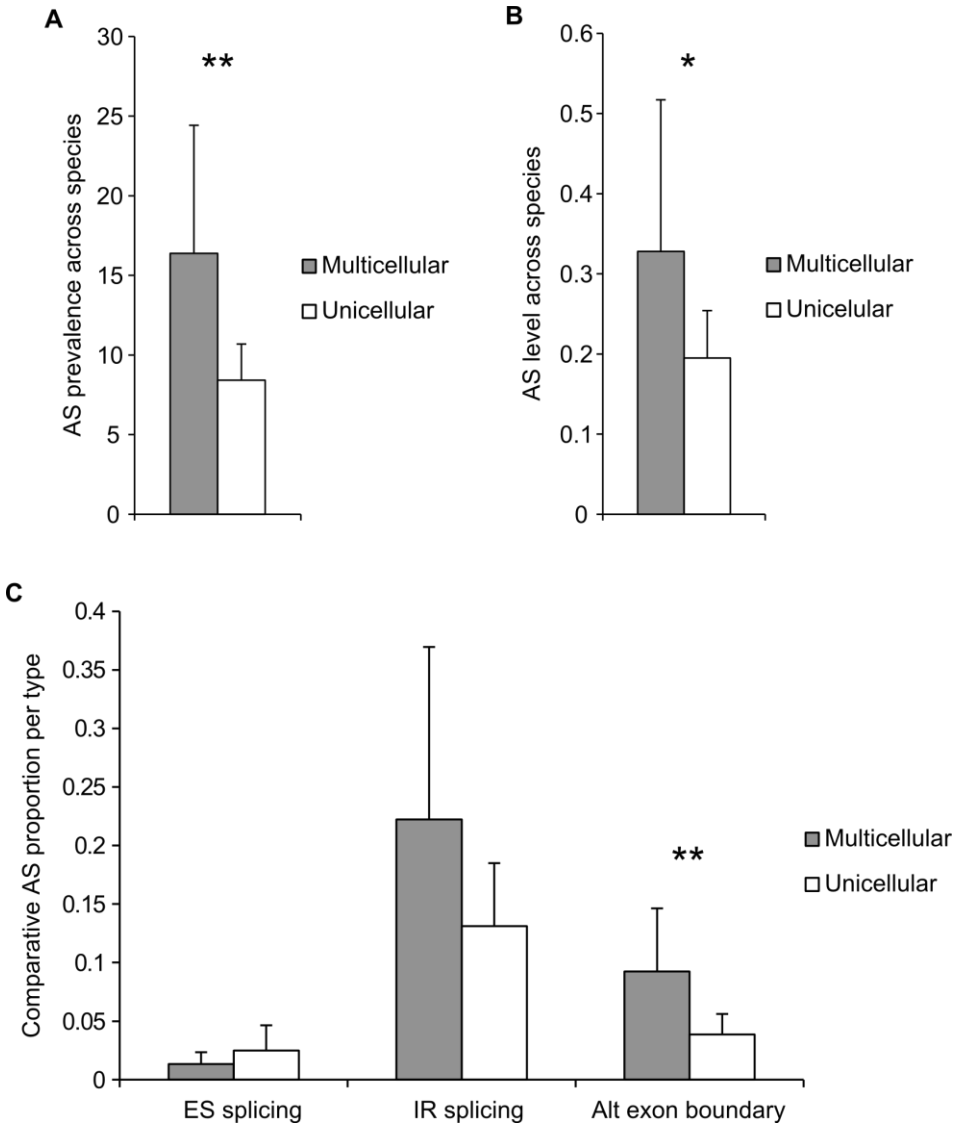


Figure 3.5. AS distribution across multicellular and unicellular fungal species. Barplots of the mean Alternative Splicing prevalence (A), mean Alternative Splicing level (B) and proportion of AS by type (C) in multicellular and unicellular species. Error bars represent standard deviation. Significance of the difference of the means was obtained through a Welch's t test (* $p < 0.05$, ** $p < 0.01$).

In metazoans, ES events have been found to be more likely to be conserved between species as compared to IR events, which may possibly reflect ES has a higher functional contribution to the transcriptome (Ast 2004). Thus, we examined

whether variations in the proportion of ES and IR AS events across species were associated with multicellularity in fungi species. We found no significant difference between the proportions of ES or IR AS events among multicellular species when compared to unicellular species (Figure 3.5, Supplementary table 3.5). No significant differences in the proportion of genes with exons skipping when comparing multi cellular and single celled species. These results suggest that although ES has been associated with conserved protein coding isoforms in metazoan species, this is a rare AS type of event among fungi species and might have only a limited impact on the overall proteome pool. Interestingly, despite being the most common event type, levels of intron retention events were no different in single *vs.* multi cellular species.

To further explore the role of AS prevalence in multicellularity, we examined transcript data from *U. maydis* species, obtained during unicellular, haploid stage (2336 ESTs) and multicellular, filamentous condition (2928 ESTs). Because of the reduced sample size we only compared the proportion of transcripts with at least one alternatively splicing event in each condition. Contrary to our expectations, we found that the proportion of AS event containing transcripts was higher in the unicellular state compared to the multicellular stage (Single celled: 3.7243%, Multi-celled: 1.9808%; two proportion Z-test, Z=3.8398, p-value = 0.00012). This result suggests that although alternative splicing might be higher among multicellular fungi species; additional isoforms in the expanded transcriptome might not be preferentially expressed in the multicellular states of individual species life cycle.

Taken together, these results are in favour of an association between AS and multicellularity in the fungal species examined which is not explained by a phylogenetic signal artefact. Our findings lend support to the suggestions for a key role for transcript diversification through alternative splicing in the evolution of increased cell type numbers in eukaryotic species (Nilsen & Graveley 2010). We found that the use of alternative exon boundaries explains the differences observed and not intron retention events despite being the most common type of event. With the continuing accumulation of transcriptome data for an ever increasing number of species, future studies should provide a fuller understanding on the functional importance of alternative splicing among fungi species.

3.4 Conclusions

Here we have characterised AS in 23 fungi species. Using comparable AS estimates for 19 of these species with the highest EST coverage per gene, we were able to compare AS prevalence across species. We found that, on average, close to 20% of studied genes are alternatively spliced. Such levels of splicing contrast with first reports for lower AS levels in fungal species. Consistent with previous studies characterising AS in fungal species we found that IR is by far the most common type of AS event.

We found high levels of variability in AS prevalence and levels overall and by AS type among the species studied. Interestingly, we found a significant association between higher alternative splicing and multicellularity consistent, in principle, with past suggestions for a role of transcript diversification in the evolution of multicellularity (Nilsen & Graveley 2010). This association was not a by-product of phylogenetic relatedness.

To the best of our knowledge this study represents the most extensive characterisation of AS prevalence in a fungi species including species from three different phylogenetic groups of fungi (*Ascomycota*, *Basidiomycota* and *Zygomycota*). Moreover, this is the first analysis to provide AS indexes controlling for differences in transcript coverage allowing the comparison of prevalence of AS among 19 fungi species. This study includes species from three different phylogenetic groups of fungi (Ascomycota, Basidiomycota and Zygomycota) but with a larger number of species than previous studies (McGuire et al. 2008) providing a broader phylogenetic perspective. Importantly, we present the first assessment of AS levels in a number of species: A. niger, C. heterostrophus, C. parasitica, G. moniliformis, G. zeae, M. fijiensis, M. graminícola, N. haematococca, T. terrestris, T. atroviride, M. laricis-populina, S. roseus, M. circinelloides, P. blakesleeanus.

3.5 Supplementary tables

Supplementary table 3.1. Transcript and genome annotations data sources.

Table shows sources and versions of genomic data used for studies. For each species it also shows the number of transcripts used to identify AS events.

Species	Genome version	Number of transcripts	Source	Annotation
Aspergillus flavus	Afl_CADRE.1	20371	ftp://ftp.ensembl.org/pub/release-62/fasta/	Ensembl
Aspergillus niger	Ani_CADRE.1	46938	ftp://ftp.ensembl.org/pub/release-62/fasta/	Ensembl
Aspergillus terreus	Ate_CADRE.1	12776	ftp://ftp.ensembl.org/pub/release-62/fasta/	Ensembl
Coccidioides immitis	Cim_rs_3	62729	ftp://ftp.jgi-psf.org/pub/JGI_data/	Broadinstitute
Coccidioides posadasii	Cpo_rmscc_3488	110163	ftp://ftp.ncbi.nih.gov/genomes/	Broadinstitute
Cochliobolus heterostrophus C5 (Bipolaris maydis)	CocheC5_1	28747	ftp://ftp.jgi-psf.org/pub/JGI_data/	JGI
Cryphonectria parasitica	Cpar_v2	41858	ftp://ftp.jgi-psf.org/pub/JGI_data/	JGI
Gibberella moniliformis (Fusarium verticillioides)	Gmo_FV3.3	87086	ftp://ftp.ensembl.org/pub/release-62/fasta/	Ensembl
Gibberella zeae (Fusarium graminearum)	Gze_FG3.3	21355	ftp://ftp.ensembl.org/pub/release-62/fasta/	Ensembl
Leptosphaeria maculans (Phoma lingam)	Lepmu1	42319	ftp://ftp.jgi-psf.org/pub/JGI_data/	JGI
Melampsora laricis-populina	Mpo_JGIv1	54445	ftp://ftp.ensembl.org/pub/release-62/fasta/	JGI
Mucor circinelloides (Rhizomucor circinelloides)	Mci_JGIv2	27614	ftp://ftp.jgi-psf.org/pub/JGI_data/	JGI
Mycosphaerella fijiensis (Paracercospora fijiensis)	Mfi_JGIv2	36233	ftp://ftp.jgi-psf.org/pub/JGI_data/	JGI
Mycosphaerella graminicola (Septoria tritici)	Mgr_JGIv2	32194	ftp://ftp.jgi-psf.org/pub/JGI_data/	JGI
Nectria haematococca MPVI	Necha2	33120	ftp://ftp.jgi-psf.org/pub/JGI_data/	JGI
Neurospora crassa (Chrysonilia crassa)	Ncr_EF1	277147	ftp://ftp.ncbi.nih.gov/genomes/	NCBI
Phycomyces blakesleeanus	Pbl_JGIv2	47847	ftp://ftp.jgi-psf.org/pub/JGI_data/	JGI
Saccharomyces cerevisiae (Candida robusta)	Sce_uid128	34915	ftp://ftp.ncbi.nih.gov/genomes/	Ensembl

Species	Genome version	Number of transcripts	Source	Annotation
Schizosaccharomyces pombe	Spo_EF1	109202	ftp://ftp.ncbi.nih.gov/genomes/	NCBI
Sporobolomyces roseus	Sro_JGIv1	27048	ftp://ftp.jgi-psf.org/pub/JGI_data/	JGI
Thielavia terrestris	Thite2	27991	ftp://ftp.jgi-psf.org/pub/JGI_data/	JGI
Trichoderma atroviride	Tat_JGIv1	35125	ftp://ftp.jgi-psf.org/pub/JGI_data/	JGI
Ustilago maydis	Uma_JGI	39308	ftp://ftp.jgi-psf.org/pub/JGI_data/	JGI
Arabidopsis thaliana	AT_TAIR9.171	1529700	ftp://ftp.plantgdb.org/download/Genomes/	PlantGDB
Caenorhabditis elegans	Cel_WS220.62	396687	ftp://ftp.ensembl.org/pub/release-62/fasta/	Ensembl
Drossophila melanogaster	Dm_BDGP5.25.62	821005	ftp://ftp.ensembl.org/pub/release-62/fasta/	Ensembl
Homo sapiens	Hs_GRCh37.62Chr	8315122	ftp://ftp.ensembl.org/pub/release-62/fasta/	Ensembl
Mus musculus	Mmu_NCBI37.62	4853546	ftp://ftp.ensembl.org/pub/release-62/fasta/	Ensembl
Oryza sativa	OS_V6.1	1252989	ftp://ftp.plantgdb.org/download/Genomes/	PlantGDB

Supplementary table 3.2 Non-comparative AS statistics.

Table contains the EST coverage and total number of AS events detected per species. It also shows the non-comparative distribution by AS event type. Events where divided in three groups: exon skipping (ES), intron retention (IR) and alternative exon boundary (AEB) with all the events associated with alternative 3' and 5' splice sites.

Species	ESTs mapping to annotated genes	Total number of AS events	Distribution by type		
			AEB	ES	IR
Aspergillus flavus	6104	22	18.18%	4.55%	77.27%
Aspergillus niger	9042	26	26.92%	11.54%	61.54%
Aspergillus terreus	1331	5	20.00%	0.00%	80.00%
Coccidioides immitis	33671	583	27.27%	1.03%	71.70%
Coccidioides posadasii	58828	1606	19.12%	1.99%	78.89%
Cochliobolus heterostrophus C5	19643	59	28.81%	1.69%	69.49%
Cryphonectria parasitica	25146	192	28.13%	7.29%	64.58%
Gibberella moniliformis	62276	406	20.69%	4.43%	74.88%
Gibberella zeae	14064	120	30.83%	7.50%	61.67%
Leptosphaeria maculans	2926	12	33.33%	0.00%	66.67%
Mycosphaerella fijiensis	18306	77	29.87%	5.19%	64.94%
Mycosphaerella graminicola	14489	113	30.09%	4.42%	65.49%
Nectria haematococca MPVI	19997	92	26.09%	5.43%	68.48%
Neurospora crassa	125829	1210	27.02%	3.88%	69.09%
Saccharomyces cerevisiae	1335	1	0.00%	0.00%	100.00%
Schizosaccharomyces pombe	8238	15	20.00%	13.33%	66.67%
Thielavia terrestris	16787	113	16.81%	2.65%	80.53%
Trichoderma atroviride	22235	257	21.01%	4.67%	74.32%
Melampsora laricis-populina	37281	476	45.17%	9.45%	45.38%
Sporobolomyces roseus	9126	34	26.47%	23.53%	50.00%
Ustilago maydis	10956	24	62.50%	0.00%	37.50%
Mucor circinelloides	17861	41	31.71%	9.76%	58.54%
Phycomyces blakesleeanus	31695	130	37.69%	4.62%	57.69%

Supplementary table 3.3. Comparative prevalence of AS.

Table with details for comparative AS data for fungi species. The 19 species marked in the first column are those with a minimum of 50 genes with comparable data, those are the species used for all comparative studies. The AS prevalence is the percentage of genes associated with AS. The average AS levels were obtained by averaging the AS levels in those genes with comparable data. The last three columns show the comparative levels of AS per type, using the addition of AS levels by AS event type. Event types where divided in three groups: exon skipping (ES), intron retention (IR) and alternative exon boundary (AEB) with all the events associated with alternative 3' and 5' splice sites.

Species	Gene number *	AS prevalence	Average AS levels	AS levels by AS event type		
				AEB	ES	IR
A. flavus	47	25.53%	0.416	3.5300	0.7700	15.2400
A. niger **	98	12.24%	0.21	5.3700	2.6700	12.5600
A. terreus	10	0.00%	0.332	0.7900	0.0000	2.5300
C. immitis **	1078	20.32%	0.45	126.1405	4.3300	354.1995
C. posadasii **	1635	33.82%	0.787	239.9490	22.8108	1024.5502
C. heterostrophus C5 **	268	10.45%	0.184	14.0642	0.5725	34.5933
C. parasitica **	488	11.27%	0.266	37.5152	8.4843	83.6205
G. moniliformis **	1392	13.15%	0.238	65.0570	9.8067	256.6263
G. zeae **	276	25.36%	0.371	31.2000	6.3150	64.7750
L. maculans	16	18.75%	0.388	2.1100	0.0000	4.0900
M. fijiensis **	323	10.22%	0.197	17.8150	2.8050	42.8700
M. graminicola **	243	21.40%	0.361	25.1000	3.4367	59.1533
N. haematococca MPVI **	327	11.01%	0.225	17.2800	2.9400	53.2400
N. crassa **	1109	22.00%	0.475	147.3316	17.3410	361.9274
S. cerevisiae	3	0.00%	0.17	0.0000	0.0000	0.5100
S. pombe **	108	6.48%	0.127	2.4800	2.0000	9.2100
T. terrestris **	319	7.84%	0.233	11.3650	2.3800	60.6950
T. atroviride **	410	21.71%	0.51	40.6883	6.5700	161.9617
M. laricis-populina **	546	26.19%	0.577	136.7674	21.6587	156.6940
S. roseus **	119	10.92%	0.225	6.8383	5.8217	14.0700
U. maydis **	163	7.36%	0.129	13.3900	0.0000	7.6400
M. circinelloides **	328	6.40%	0.103	11.1533	2.6100	19.9567
P. blakesleeanus **	660	9.09%	0.166	40.9550	4.2400	64.5150

* Total number of genes with enough ESTs to use randomized sampling.
** Species used for comparative AS analyses.

Supplementary table 3.4. Phenotypic characteristic in fungi species.

Relationship of species and associated phenotype used for analyses. References included for species where different phenotypes have been reported in literature.

Species	Multicellularity	Reference
A. niger	Multicellular	(Krijgsheld et al. 2013)
C. immitis	Multicellular	(Sun & Huppert 1976)
C. posadasii	Multicellular	(Dionne et al. 2006)
C. heterostrophus	Multicellular	
C. parasitica	Multicellular	
G. moniliformis	Multicellular	
G. zeae	Multicellular	(Lee et al. 2003)
M. laricis-populina	Multicellular	
M. circinelloides	Multicellular	(Li et al. 2011)
M. fijiensis	Multicellular	
M. graminicola	Multicellular	
N. haematococca MPVI	Multicellular	
N. crassa	Multicellular	(Galagan et al. 2003)
P. blakesleeanus	Multicellular	(Fischer-Parton et al. 2000)
S. pombe	Unicellular	(Yanagida 2002)
S. roseus	Unicellular	
T. terrestris	Unicellular	
T. atroviride	Multicellular	
U. maydis	Multicellular	

Supplementary table 3.5. Relationship between AS and multicellularity.

Table with summarised results for statistical tests used to identify significant differences in AS properties between multicellular and unicellular fungi. AS event types were organized in three groups, exon skipping (ES), intron retention (IR) and alternative exon boundary (alternative 3' or 5' splice sites).

AS Property	Multicellular/ Unicellular (Welch T)	p-value
AS prevalence (% genes with AS)	3.3134	0.005646
AS level (average number of AS events per gene)	2.2808	0.04266
ES splicing proportion	-0.925	0.4459
IR splicing proportion	1.8894	0.09074
ES/(ES+IR)	-1.4389	0.2812
Alternative exon boundary proportion	3.1993	0.00857

4 Alternative lice have alternative splice

4.1 Introduction

Sucking lice are obligate hematophagous ectoparasites of placental mammals with direct life cycles, meaning they can only infect a single host species (Durden & Musser 1994). This high degree of specialism and host dependence leads to long-term co-evolution, such that speciation events of the host and parasite are congruent (Paterson & Gray 1997). Humans represent a special case as they are parasitized by three different types of lice, each of them colonising a specific region of the body (head, body and pubic area) (Reed et al. 2007; Weiss 2009). The association between the louse *Pediculus humanus* and the human host is a very ancient one, extending back at least 6 million years to the last common ancestor of humans and chimpanzees. *P. pubis* are thought to have descended from the *Pthirus* which parasites gorillas (Reed et al. 2007), colonising humans at some point after they lost most of their corporal hair and *P. humanus* became restricted to the head area. *Pediculus humanus* has been sub-classified into *P. humanus capitis* (head lice) and *P. humanus humanus* (body lice) (Light et al. 2008). Whereas head lice correspond to the ancestral lineage, body lice have derived from head lice relatively recently, and probably on multiple occasions (Li et al. 2010). As the female body louse lays eggs exclusively on the host's clothes (Light et al. 2008) it is thought that it emerged from head lice after the use of clothing became widespread approximately 170000 years ago (Kittler et al. 2003; Toups et al. 2011).

Head and body lice were originally classified as distinct species in the mid-18th century on the basis of morphological and ecological traits (De Geer 1767). However, this distinction is highly contentious (Buxton 1947; Nuttall 1940). Whilst the two groups are not known to interbreed in the wild (Schaefer 1978; Busvine 1948), they can produce fertile offspring under experimental conditions (Bacot 2009; Mullen & Durden 2002). Furthermore, although microsatellite data has been used to

argue that head and body lice co-infecting a single human host do indeed represent distinct species (Leo et al. 2005), most of the molecular data available do not support this view. For example, a comparison of ten cytochrome oxidase I (COI) gene haplotypes representing both groups pointed to a single species (Leo et al. 2002), and this conclusion was supported by subsequent analysis of six gene loci (Light et al. 2008). Phylogenetic analysis of global samples of head and body lice, based on 18s rRNA (Leo & Barker 2005), COI and cytB (Reed et al. 2004) and intergenic spacers (Veracx et al. 2012; Li et al. 2010), all support the view that the two groups are not phylogenetically distinct or monophyletic, and provide evidence for ongoing gene-flow between them. Thus, despite the fact that body lice have distinct morphological (see Figure 4.1 for sample differences) and ecological characters compared to head lice, there remains little justification for considering head and body lice separate species.

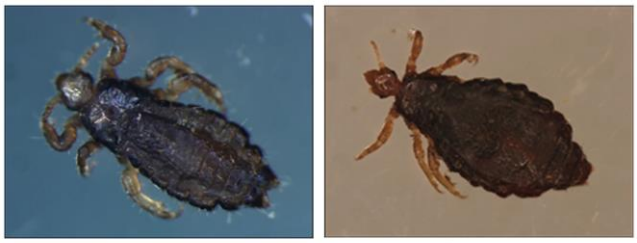


Figure 4.1. Showing different morphologies of Clade A head louse and Clade A body louse. Figure adapted from (Veracx et al. 2012).

Regardless of the taxonomic debates, the different niches occupied by head lice and body lice have significant relevance for public health. Whereas head lice are very common, particularly in children, they are not known to be competent vector for any infectious agent. In contrast, body lice are more rarely encountered and tend to infect adults living in very poor sanitary conditions, such as the homeless or warring soldiers. However, body lice are competent vectors of three bacterial pathogens; *Rickettsia prowazekii* (the causative agent of epidemic typhus), *Bartonella quintana* (trench fever) and *Borrelia recurrentis* (relapsing fever) (Rydkina et al. 1999; Weiss 2009).

The recent sequencing of the body louse genome, along with its primary endosymbiont, represent a significant advance in understanding both the ecology and

evolution of this parasite in relation to other insect species (Kirkness et al. 2010). These efforts, however, have shed little light on the bases for the shift in ecology and vector competence when comparing head and body lice. The louse genome is the smallest known insect genome; 90% of the annotated genes share homologues in other insect species (Kirkness et al. 2010; Pittendrigh et al. 2006), and EST data have confirmed that all transcript producing genes have been annotated in the initial genome release (Olds et al. 2012). It is also clear from EST data that the transcriptomes of head and body lice are essentially identical (Olds et al. 2012; Drali et al. 2013). It has been suggested that the phenotypic shifts associated with the emergence of body lice are likely to be a consequence of a small number of point mutations or may result from subtle regulatory changes, possibly epigenetic in origin, triggered by environmental cues (Li et al. 2010). Notably, variation in the PHUM540560 gene constitutes the only genetic marker identified to date which can distinguish head from body lice once they are removed from their habitat (Drali et al. 2013). Thus, the genomic features associated with the phenotypic differences between body and head lice remain unknown.

AS is a common posttranscriptional process by which multiple distinct transcripts, and hence proteins, can be encoded from a single gene by differentially splicing exons. The process provides a means for rapid functional innovations via very economical genomic changes (Chen et al. 2012; Nilsen & Graveley 2010). Novel AS regulatory elements can be generated with a few mutations, thus facilitating both short and long-term evolutionary adaptation even in the absence of changes in the gene content through gene duplication and/or gene loss (Chen et al. 2012). A recent study in *Mus musculus* subspecies suggests that using next generation sequencing data to compare the effects of AS on the transcriptome of species which recently diverged show AS is a relevant source of transcript variation in recent speciation events (Harr & Turner 2010). Moreover, previous studies in eusocial insects (Bonasio et al. 2012; Lyko et al. 2010), where very different phenotypes originate from the same genome, have found differences in AS patterns, a product of low level gene expression regulatory processes, corresponding with distinctive cast phenotypes.

Here we characterise AS patterns in human lice using RNA sequencing data [24] in an attempt to identify AS events which are specific to either body or head lice in order to gain insights into the role of alternative splicing in the phenotypic and disease vectoring differences observed.

4.2 Materials and methods

4.2.1 Identifying AS events

A total of 401578 partial sequences from *454* (Olds et al. 2012) were aligned to the body lice genome (Kirkness et al. 2010) using GMAP (Wu & Watanabe 2005). Reads which aligned to regions with no annotated genes were discarded from any further analysis. Alternative splicing events were identified according to Chen et al. (2011). In brief, *454* transcripts to genome alignments were used to generate exon-intron gene templates. Then AS events were identified by comparing alignment coordinates *454* sequences from head and body lice respectively against corresponding gene templates. Of the complete set of 401578 transcript sequences from *454*, a total of 86993 contained one or more alternative splicing events corresponding to a total of 10941 distinct events identified.

To produce comparable values of AS a random sampling method based on previous work by Chen et al. (2011) was used to correct for sequence coverage bias in both head and body lice. In summary, the corrected AS index is considered as the average number of AS events found in 100 samples of 10 randomly selected *454* sequences aligned to the lice genome.

4.2.2 Illumina confirmation of splice sites

Illumina RNA-Seq short reads were aligned to the genome and identified splice sites using GSNAP (Wu & Watanabe 2005). Coordinates of splice sites were then used to validate alternative splicing events identified from *454* sequences; exact matches were required.

4.2.3 Functional gene associations

Associations to immune functions were tested using three separate lists of genes: immune response (Lemaitre & Hoffmann 2007); phagocytosis, proteomics analysis of purified phagosomes (Stuart et al. 2007) and phagocytosis, RNAi screening during bacterial phagocytosis (Stuart et al. 2007). Gene ontology terms (GO) were assigned according to gene orthology with *Drosophila melanogaster* obtained from FlyBase (McQuilton et al. 2012). We obtained at least one biological process GO term annotation for a total of 4398 lice genes (58.7%). GO terms with less than 100 annotated lice genes were grouped together.

4.2.4 Randomisation tests

Overrepresentation of AS type or functional associations among head or body lice specific AS events and the genes containing them were tested through the use of a randomisation test. For this the proportion of head or body lice specific AS events (or their corresponding genes) associated with the feature of interest was calculated and contrasted with that of the wider pool of lice AS events and associated genes. Statistical significance for observed enrichments or impoverishments of AS types or functional associations was then established using a randomisation protocol obtaining 10000 random samples of the same number of AS events or genes as the ones being tested for enrichment, drawn from the whole population of alternatively spliced lice genes. The number of AS events or genes associated with the feature being tested among genes with head or body lice specific AS events is then compared with the mean of the number of genes pertaining to that same feature in the random samples by means of a Z-test. A Benjamini-Hochberg correction for multiple testing was then applied where needed.

4.3 Results

4.3.1 Alternative splicing in human lice

In order to characterise AS patterns in human head and body lice, we analysed *454* and short read Illumina data previously obtained from head and body lice [24] (see methods). A total of 86993 transcript sequences were found to contain one or more AS events and correspond to a total of 10941 distinct events identified. These events can be mapped to 3918 genes from out of the 10773 lice protein coding genes.

As lice has been found to have one of the smallest insect genomes (Kirkness et al. 2010), we assessed whether the contraction in genomic sequence and gene count was reflected in a lower overall proportion of genes undergoing AS compared to other invertebrate genomes. To assess how human lice AS patterns compare with those of other arthropod and invertebrate species, we obtained an index of AS using a random sampling method, which corrects for the bias caused by transcript coverage differences across species (see methods). Using this comparative index, we found that the prevalence of AS events in human lice fall within the range of AS prevalence in other arthropods (Figure 4.2; Supplementary table 4.1).

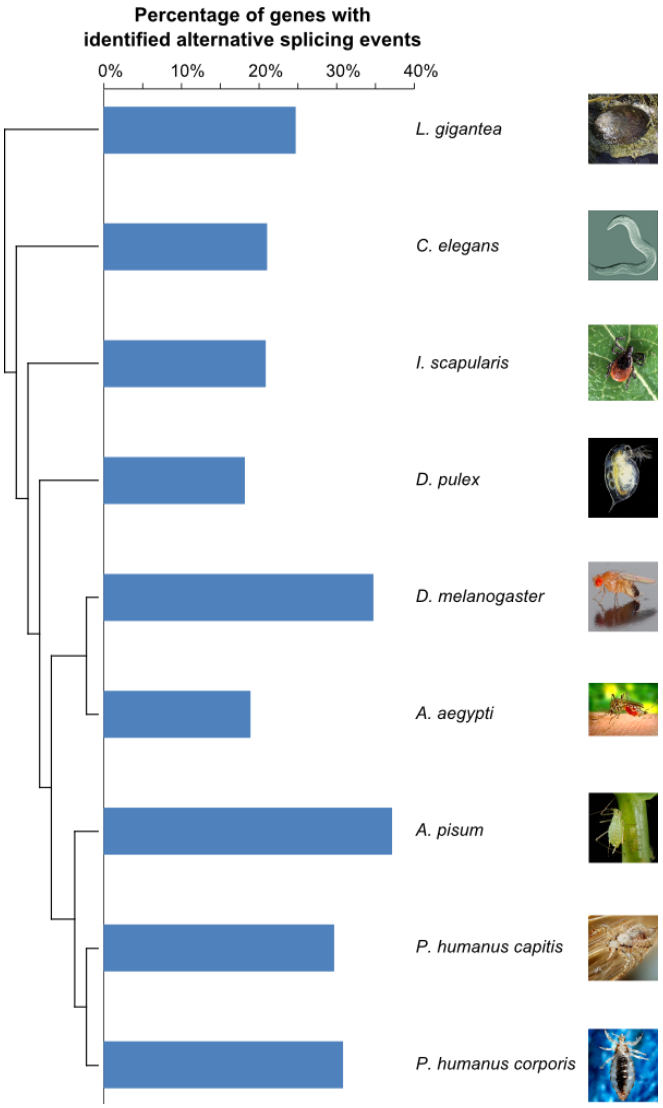


Figure 4.2. Comparison of AS prevalence in human lice with other arthropod species. Bars adjacent to phylogenetic tree represent the proportion alternatively spliced genes in different species. See methods for details of the comparative AS index. (*C. elegans* and *L. Gigantea* are included as out-group reference). See Supplementary table 4.2 for sources of images.

4.3.2 Head and body lice specific alternative splicing events

Of particular interest are events unique to either head or body lice as differences in AS might shed some light on the possible molecular bases of recent adaptations allowing human lice to survive on clothing and to vector human disease. In order to explore head or body specific AS events, we first used Illumina short reads obtained from head and body lice (Olds et al. 2012) to confirm AS events

detected from *454* sequencing (see methods). This additional data allowed us to confirm a total of 8369 (76%) of AS events identified from *454* sequences alone, 6142 in head and 6839 in body lice mapping back to 3506 lice genes. A total of 3598 of Illumina supported AS events were found to be unique to either head or body lice with 1415 and 2183 AS events respectively.

A previous study in metazoans suggested that there were variations in the functional relevance of different AS types, as events involving alternative use of exons (exon skipping, ES) were more likely to be conserved between species compared to intron retention events (Ast 2004). Thus, we examined the distribution of types among head and body specific AS events. Using a randomisation protocol (see methods), we found that both head and body lice specific AS events were significantly enriched in ES and alternative 3' acceptor site (3S) AS types compared to the total pool of lice AS events ($p = 0.0029$ for ES in head lice, $p < 0.001$ for ES in body lice and $p < 0.001$ for 3S in both head and body lice, after Bonferroni corrections; Figure 4.3 and Supplementary tables 4.3 and 4.4). We also found a significant underrepresentation of intron retention (IR), alternative 5' donor site (5S) and alternative 5' donor-3' acceptor splice sites (3S5S) for both head and body lice ($p < 0.001$ for IR and 5S in both head and body lice, $p = 0.0770$ for 3S5S in head lice and $p = 0.002$ for 3S5S in body lice, after Bonferroni corrections; Figure 4.3 and Supplementary tables 4.3 and 4.4). These findings suggest that head and body lice specific AS events are likely to be contributing to the functional transcript pool rather than represent primarily AS noise.

We then compared the overrepresentation of ES and depletion of IR in head and body lice specific AS events in order to further explore the potential functional relevance of ecotype specific AS events. Body lice specific AS events were found to have a stronger bias towards ES and depletion of IR compared to head lice (2 Prop Z for ES [body *vs.* head] = 3.692, $p = 0.0002$, 2 Prop Z for IR [body *vs.* head] = -2.0192, $p = 0.04338$; Figure 4.3). This result suggests that although both head and body lice have a significant bias in favour of ES and depletion of IR AS event type, probably reflecting that AS specific events in both body and head lice tend to functionally contribute to the transcript pool, this pattern is stronger in body lice.

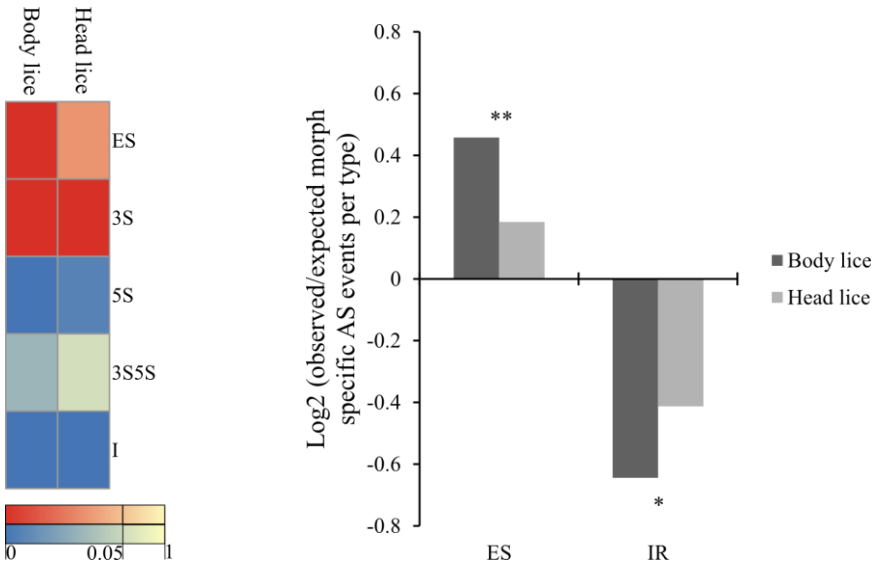


Figure 4.3. Enrichment and impoverishment of AS event types among head or body lice specific confirmed AS events. Top left panel shows heat map graph of enrichment analysis of head and body lice specific AS events compared to the wider pool of lice AS events. Colours represent enrichment (red) and depletion (blue) for each AS event type in head and body lice. Higher colour intensity reflects the statistical significance of enrichment and depletion of each AS event type with yellow tones for less significant deviations from random expectation (p-values after Bonferroni correction). The figure shows 5 types of AS events studies, exon skipping (ES), alternative 3' splice site (3S), alternative 5' splice site (5S) alternative 3' and 5' splice sites (3S5S) and intron retention (I). All AS types are significantly deviated from random expectations in both head and body lice except for 3S5S AS events. Top right panel shows significant enrichment of exon skipping and significant depletion of intron retention in both head and body lice. Each pair of columns represent the proportion of over/under representation of each type as compared with the expected number of events detected for each specific AS type on each lice type. Stars represent significance of the result (*) for p< 0.05 and (**) for p<0.005.

4.3.3 Enrichment of functional associations among body lice specific AS events

To assess the potential functional relevance of the inferred head or body lice specific AS events, we examined functional associations of the sets of genes affected. The 3598 head or body lice specific AS events mapped to a total of 2016 louse genes, each containing one or more AS event only found in either head or body lice derived transcripts. Of these genes, 974 genes contained AS events unique to head lice and 1309 contained at least one AS event specific to body lice. A minority of genes (267 genes) lie in the interception, meaning the same gene was found to be associated to both a head louse specific and body louse specific AS event (Figure 4.4).

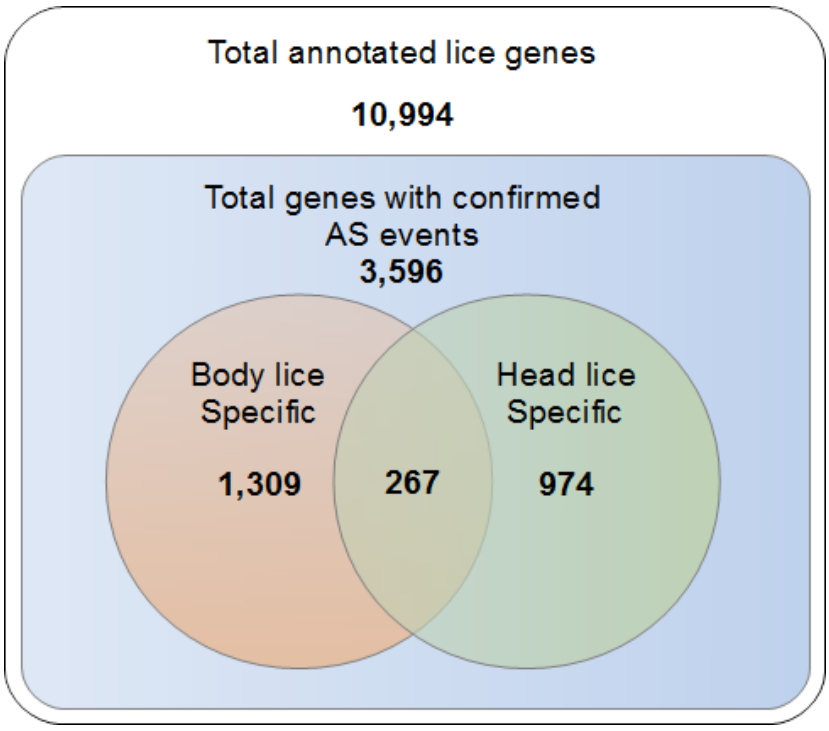


Figure 4.4. Diagram of alternatively spliced gene counts in head and body lice. The most outer square shows the total number of annotated lice genes. The inner square represents those genes with at least one confirmed AS event in either head or body lice. The left circle includes genes with at least one AS event unique to body lice. The right set includes genes with at least one AS event unique in head lice, the intersection includes genes with both at least one unique AS event in head and one unique AS event in body lice.

As immune system genes and reproductive genes have been previously associated with processes of breeding incompatibility during speciation processes (reviewed in (Eizaguirre et al. 2009; Brucker & Bordenstein 2012; Johnson 2010)), we explored whether there was any evidence for the enrichment of these genes among those with head or body lice specific AS events. In addition, immune system response genes have been found to be associated with vector competence in the mosquito (Smith et al. 2011).

We tested for enrichment among genes with head or body lice specific AS events using three independently compiled lists of immune related genes (see methods). Using a randomisation test, we found no evidence of a significant enrichment of immune related genes among genes with either head or body lice specific AS events when compared to the wider pool of alternatively spliced genes. This finding was the case in all three lists tested even if multiple test correction is forfeited ($p > 0.05$). We then explored the possibility that representation of reproductive function associated genes were observed among those with AS events specific to head or body lice. For this, we used a previously compiled list of genes with reproductive function (Wasbrough et al. 2010). We found a significant enrichment of reproductive genes among those with body lice specific AS but not with head lice specific AS events (genes with head lice specific AS $p = 0.1441536$; genes body lice specific AS $p = 0.01185075$).

These results show that there is a significant enrichment of reproductive function associated genes among those with AS events specific to body lice. No such enrichments were observed in any immune related gene lists from either louse. While enrichment of reproductive associated genes among genes containing body lice specific AS events is suggestive of possible incipient reproductive isolation, this finding may very well reflect selective pressures acting on the body lice phenotype to life in human clothing.

To gain further insights into the functions of genes associated with head or body lice specific AS events, we assigned gene ontology (GO) terms to lice genes according to annotations in corresponding Drosophila melanogaster orthologous

genes (see methods). We ascribed at least one GO term annotation for a total of 4398 lice genes (58.7% of the total gene pool). We then examined whether any functional categories were significantly overrepresented among genes with head and body specific AS events compared to the wider pool of alternatively spliced lice genes (see methods). No GO terms were found to be significantly enriched when assessing genes with AS events unique to head lice. In contrast, genes affected by AS events unique to body lice were significantly associated with peripheral nervous system development, salivary gland morphogenesis, open tracheal system development, dorsal closure, regulation of transcription (DNA dependent), ovarian follicle cell development and autophagic cell death (Figure 4.5 and Supplementary table 4.5). These results show that while head lice specific AS events affect genes representative of the wider pool of alternatively spliced genes, those in body lice show significant deviations in favour of genes associated with neuronal connectivity and development.

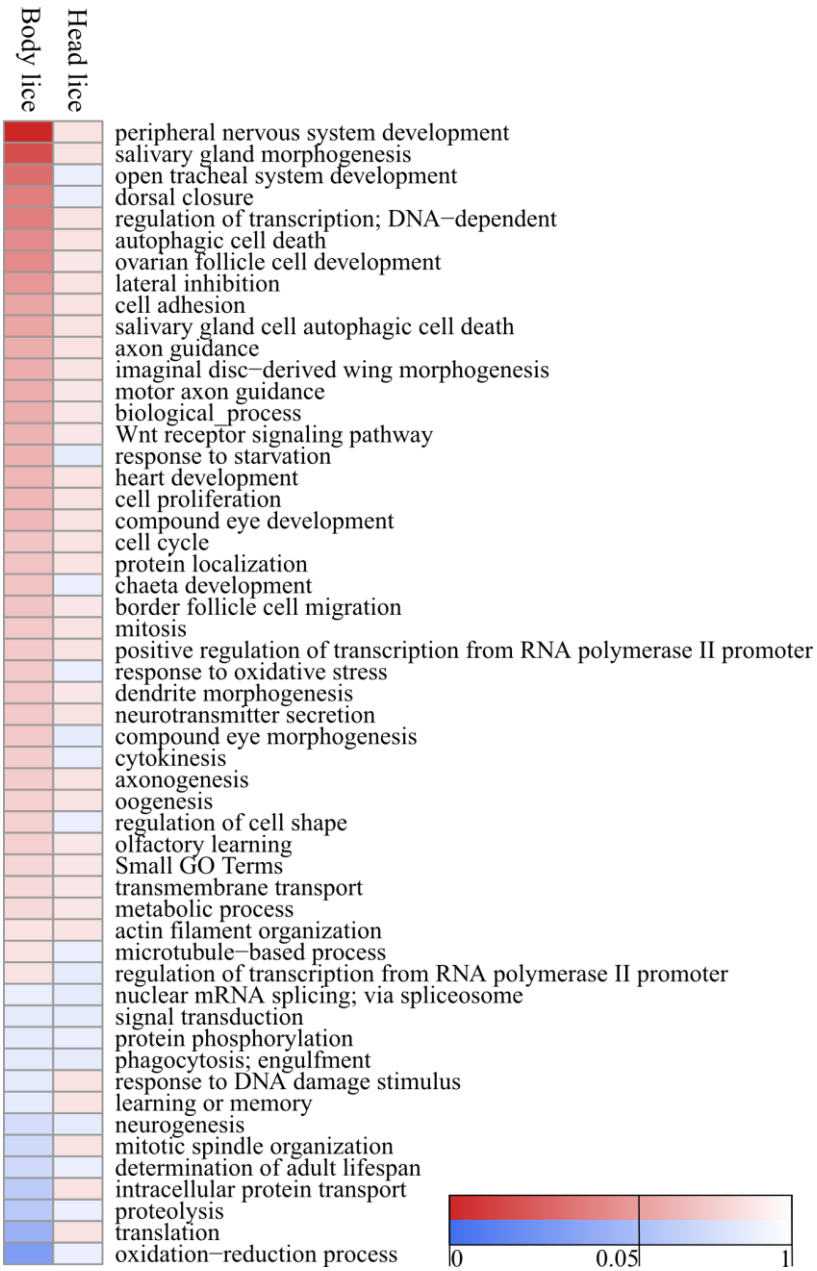


Figure 4.5. Gene ontology enrichment among genes with unique alternative splicing events in head or body lice compared to the wider set of alternatively spliced lice genes in either lice. Colours, red and blue, represent enrichment and depletion respectively. Darker shades represent stronger statistical support for the observed variation. GO categories with fewer than 100 genes were grouped together. Side bar shows the significance threshold for overrepresentation and underrepresentation of GO categories after Benjamini-Hochberg correction.

4.4 Discussion

Here we have characterised AS patterns in human head and body lice by analysing previously reported transcriptomes for head and body lice (Olds et al. 2012). We estimate that around 30% of lice genes are alternatively spliced. Such proportion is consistent with our results for levels of AS found in other insects using comparable methods. These results are of special interest because even though the human head and body lice share a compact genome and is expected for parasites to stream-line their genomes, similar AS in free-living and these obligate parasites initially suggests prevalence of AS may not follow the same pattern of austerity. Previous studies in unicellular (Merhej et al. 2009) and multicellular (Opperman et al. 2008) organisms show obligate parasites are subject to evolutionary forces which favour compact genomes. Factors like a predictable environment, using the host as a provider of biological functions, reduced effective populations and an evolution linked with the host; may result in a significant reduction in the number of genes. But then, why the AS prevalence is not dramatically reduced? If conservation of protein family diversity is favoured in contrast with the loss of genetic redundancy in obligate parasites (Mendonça et al. 2011), then one possible explanation is for AS to be compensating for the loss of one parologue by providing extra functionality to the remaining copy. Although this study has big limitations in terms of the number of species, the impossibility to differentiate data from different developmental stages, a limited number of individuals sampled and no experimental support for the functionality of the AS events identified, still is interesting to consider the possibility of AS working as a repository of functions in compact parasite genomes.

Although we were able to identify over 3000 AS events specific to only one type of louse and interestingly body lice presented a higher proportion of AS events not found in head lice than vice versa involving a higher number of genes. We are aware there is a limitation on the conclusions we can draw from these results as there are no replicates. Since sequences of all individuals of each species where treated as a group, is impossible to define if AS events specific to one type of louse are present in many or only one of the individuals sampled of each group. Rising the possibility

of lice type specific AS events being transcription or sequencing errors rather than functional transcripts.

AS has potential to diversify an organism's proteome, and changes in AS patterns have been associated with different stages of organism development (Venables et al. 2012), sex determination (Salz 2011; Venables et al. 2012), and with neural and immune system processes (Kurtz & Armitage 2006; Venables et al. 2012; Watson et al. 2005). More importantly, AS has been found to be strongly associated with transcript variation in early speciation events during a study involving mouse subspecies (Harr & Turner 2010) and also with the phenotypic plasticity in eusocial insects (Bonasio et al. 2012; Lyko et al. 2010), where distinct phenotypes derive from a single genome. Thus, AS events in the transcriptome of body lice could partly reflect the impact of adaptive evolution on the transcriptional patterns of body lice, as it occupies a novel environment to that of head lice, irrespective whether this louse represents an early speciation or an alternative head louse phenotype. It could also derive from defects in the splicing machinery, however, generating spurious spliced variants possibly resulting from the increased stress body lice endures compared to head lice. Indeed, past studies have shown that a proportion of AS events identified in the human and mouse transcriptome are likely to be non-functional and result from splicing errors (Sorek et al. 2004). Also the impossibility to pair AS events detected with specific developmental stages greatly reduces the scope of the conclusions, as it is impossible to ascertain lice type specific AS events detected are directly affecting the morphologies of the organisms in early developmental stages.

In order to explore the likelihood of the scenario of AS facilitating adaptation, we proceeded to characterise the head and body lice specific AS events by comparing them to the wider pool of events found in either louse. We first examined the distribution of head and body lice specific AS events by type. We found a significant overrepresentation of 3S and ES types in both head and body lice specific AS events. We also observed an underrepresentation of 5S and IR types in both of these lice. ES is the AS type considered to be most likely associated with functional genomic innovation and a potential driver of functional complexity in higher eukaryotes (Keren et al. 2010). In contrast, IR splicing events have been found to be

more commonly associated with lower conservation between species (Ast 2004) and possibly resulting in fewer functional transcripts and lower impact on phenotype complexity (Kim et al. 2008a). The fact that these enrichments/impoverishments were unique or stronger in body lice transcriptomes compared to head lice transcriptomes is consistent with an adaptive scenario specific to the body lice transcriptome. Nevertheless, is impossible to confirm this results without support from experimental data. Moreover, even if AS is taking part in the adaptive changes of the organisms, is impossible to ascertain adaptive changes will be present even without the presence of AS, in other words we cannot exclusively link the divergence of these two types of lice with AS because there is no experimental control which removes head lice type specific AS events and directly results in the incapability to produce viable head lice.

To gain further insights into the potential functional relevance of AS events, we then examined the functional associations of genes with AS events specific to head or body lice. If head or body lice specific AS events are the result of neutral variations in alternative splicing patterns or selective pressures not related to the specific challenges associated with each phenotype, then we would expect that the functional annotations of genes with head or body specific AS events would reflect that of the general population of alternatively spliced genes. Changes in the regulation of immune and reproductive associated genes could contribute to a scenario of early speciation by reducing gene flow by boosting fertility incompatibilities (Eizaguirre et al. 2009; Brucker & Bordenstein 2012; Johnson 2010). Indeed, AS in immune related genes has been found to be an important factor for the interaction between parasitic microorganisms and their insect vectors (Smith et al. 2011; Dong et al. 2006). We found, no significant enrichment among genes with AS events specific to either head or body lice related to immune system response after evaluating three independently compiled lists of immune related genes (Stuart et al. 2007; Lemaitre & Hoffmann 2007).

A significant enrichment of reproductive associated genes was observed among genes with body lice specific AS events. It is possible that this enrichment might reflect selective pressures favouring increasing genetic isolation between head

and body lice. However, it could also result from the particular challenges faced by human lice to maximise the fitness of the body lice phenotype in the absence of a speciation process. Lice are known to depend on symbiotic relationships with bacteria in order to survive solely on a human blood diet (Sasaki-Fukatsu et al. 2006). The primary bacterial endosymbiont, *Riesia pediculicola*, has been repeatedly confirmed to migrate in female lice and infect the ovarian epithelium, passing them to the next generation of lice (Sasaki-Fukatsu et al. 2006). Thus, the observed enrichment in reproductive genes could result from the nutritional requirements of lice to survive in clothing. Nevertheless, further studies are needed in order to confirm this possibility. Sex related genes have been observed to be subject to rapid evolution in other closely related insect species (Haerty et al. 2007) and the differences in AS in head and body lice may just show different selection pressure acting in these two types of lice.

Analyses of the representation of GO terms revealed that genes with body lice specific AS events are enriched in peripheral nervous system development, salivary gland morphogenesis, open tracheal system development, regulation of transcription <DNA-dependant>, dorsal closure, autophagic cell death and ovarian follicle cell development. Notably, no category being overrepresented among genes with head lice specific AS events. Interestingly, although the transcriptome data was obtained from pooled samples of head and body lice at different stages in their life cycles (Olds et al. 2012), all of the enriched functional categories are related to development, suggesting that some of the transcripts unique to body lice may underpin some of the behavioural and morphological differences between the two lice.

Changes in ovarian follicle cell development of body lice are also consistent with an enrichment in reproductive genes in body lice specific AS events and may reflect changes in the body louse's relationship with its endosymbiont, which must supply the larger body louse with more B vitamins (Sasaki-Fukatsu et al. 2006).

Several genes with nervous system development functions have been implicated with cast differentiation in the honey bee (Barchuk et al. 2007). The

significant enrichment of peripheral nervous system development functional associations among genes with body lice specific AS events suggests a possible role for these alternative transcripts in the behavioural or sensory function differences between these two lice. Terms related to other aspects of nervous system connectivity and development were also found to be enriched (not significant after multiple test correction) among genes with body lice specific AS events, but not in head lice. This finding further supports the contention that AS events specific to body lice from nervous system related genes may play a role in the adaptations of body lice phenotype differences compared to head lice.

In insects and other invertebrates, autophagic programmed cell death is central to normal and stress induced development (Baehrecke 2002). In particular, this process has been shown to play an important role in early gametogenesis (Nezis et al. 2009), metamorphosis (Juhász et al. 2003; Truman et al. 1994; Liu et al. 2009; Rusten et al. 2004; Mirth & Riddiford 2007), nutrition-dependant growth rate, central nervous system structure definition (Truman et al. 1994) and remodelling of specific body structures during and between moulting (Lee & Baehrecke 2001; Juhász et al. 2003; Rusten et al. 2004). Programmed cell death associated genes have been shown to play a role in producing alternative morphologies including the dauer larvae in *C. elegans* in response to environmental stress (population density and food availability) (Liu et al. 2009), and alternative casts in the honey bee (Barchuk et al. 2007). It is thus possible that AS events specific to body lice produced by genes associated with autophagy might play a role in morphological adaptations of feeding structures and potentially reproductive organs in human lice during development of the body lice phenotype.

Interestingly, several of the enriched functions (salivary gland morphogenesis, open tracheal system development, dorsal closure and autophagic cell death) are directly related to morphogenesis further supporting a potential role in the development of alternative lice morphs for body lice specific AS events. In particular, salivary gland morphogenesis and open tracheal system development might reflect feeding adaptations in the body lice phenotype compared to head lice. Notably, the salivary gland has been related to vector competence in other insects

(Gray 1998; Bowman et al. 1997) and it is thus possible for body lice specific AS events related from genes associated with salivary gland morphogenesis may play a role in the vector competence differences apparent between these two lice.

Although all these results seem to link AS with relevant functional changes between head and body lice, we are aware there is a strong limitation on the conclusion we can draw as we fundamentally rely on the correctness of the fruit fly GO. Therefore individual results may be subject to imprecisions product of problems with the quality of the GO annotations or the misinterpretation of GO terms associated with very broad functions.

It is important to note that while significant differences in the transcript pool in head and body lice were found, these differences do not contribute to the lice species classifications debate. The observed differences in the transcript pool between head and body lice could result from a reduced number of polymorphic sites or changes in the methylation patterns of AS regulatory regions. Indeed, a link between methylation and AS patterns in relation to phenotypic differences has been suggested in the ant *Camponotus floridanus* (Bonasio et al. 2012) and in *Apis mellifera* (Jarosch et al. 2011), where methylation and resulting differences in AS patterns have been associated with the capability to produce the phenotypic plasticity required for a caste system in eusocial insects.

In conclusion, the comparison of head and body lice transcriptomes, presented in this study, has revealed differences in AS patterns. Even with the limitations this kind of study may present, these differences provide the first insights into the differences in the transcript pool which may contribute to the lifestyle and vector competence variation observed for head and body lice. Importantly, the fact that head lice specific AS events appear to largely resemble the background gene pool of alternatively spliced lice genes while significant deviations from random expectations were found among body lice genes, may reflect the unequal selective pressures relating to adaptations of human lice required for surviving in a novel environment of human clothing. The nature of the overrepresented GO terms, suggest that changes in the developmental programme, particularly in relation to

feeding and development of the nervous system may have played a role in the adaptation of lice to a clothing based lifestyle.

4.5 Supplementary Tables

Supplementary table 4.1. Compartive AS prevalence in lice and model species.

Comparison of alternative splicing prevalence in head and body lice to other invertebrate species. Human and mouse are included for reference.

Subgroup	Scientific name	Estimated AS prevalence using comparative index AS
Insects	*Pediculus humanus capitis*	0.2974
Insects	*Pediculus humanus corporis*	0.3089
Insects	*Drosophila melanogaster*	0.4166
Crustacea	*Daphnia pulex*	0.2312
Arachnida	*Ixodes scapularis*	0.2683
Roundworms	*Caenorhabditis elegans*	0.2785
Mammals	*Homo sapiens*	0.8909
Mammals	*Mus musculus*	0.842

Supplementary table 4.2. Image sources

Species name	Source
Pediculus humanus capitis	Gilles San Martin, "Male human head louse" August 17, 2010 via Flickr, Creative Commons Attribution-Share Alike http://commons.wikimedia.org/wiki/File:Male_human_head_louse.jpg
Pediculus humanus corporis	CDC/ Frank Collins, Ph.D., "Pediculus humanus var. corporis" 2006 via Public Health Image Library, public domain http://phil.cdc.gov/phil/details.asp?pid=9206
Drosophila melanogaster	André Karwath, "Drosophila melanogaster - side" July 16, 2005 via Wikipedia Commons, Creative Commons Attribution-Share Alike http://commons.wikimedia.org/wiki/File:Drosophila_melanogaster_-_side_(aka).jpg
Daphnia pulex	Paul Hebert, "Daphnia pulex" June 14, 2005 via Wikipedia Commons, Creative Commons Attribution http://commons.wikimedia.org/wiki/File:Daphnia_pulex.png
Ixodes scapularis	Scott Bauer, "Adult deer tick, Ixodes scapularis", public domain http://commons.wikimedia.org/wiki/File:Adult_deer_tick.jpg
Caenorhabditis elegans	Bob Goldstein, "Caenorhabditis elegans, adult hermaphrodite" 2007 via Wikipedia commons, Creative Commons Attribution-Share Alike http://commons.wikimedia.org/wiki/File:CelegansGoldsteinLabUNC.jpg
Aedes aegypti	James Gathany, "The yellow fever mosquito Aedes aegypti, taking a bloodmeal." 2005 via Public Health Image Library, public domain http://commons.wikimedia.org/wiki/File:Aedes_aegypti_bloodfeeding_CDC_Gathany.jpg
Acyrthosiphon pisum	Shipher Wu, "Pea aphids extracting sap from the stem and leaves of garden peas." February 2010 via Wikipedia commons, Creative Commons Attribution http://commons.wikimedia.org/wiki/File:Acyrthosiphon_pisum_(pea_aphid)-PLoS.jpg

Supplementary table 4.3. Enrichment analysis for head lice specific AS.

Enrichment and impoverishment of AS event types among genes with head lice specific AS events. Table shows all 5 AS types studied, exon skipping (ES), alternative 5' splice site (5S), alternative 3' splice site (3S), intron retention (IR) and alternative 3' and 5' splice sites (3S5S). Table shows expected and actual number of genes found for each event type. Last column shows the Bonferroni corrected p-value for each AS type.

AS event type	Head lice specific AS events per type	Expected number of head lice specific AS events from 10000 random samples	Standard Error	Z-score	One sided p-value	Bonferroni adjusted p-value
ES	398	350.3187	14.71017	3.2413	0.000594754	0.002973767
5S	303	368.1552	14.95069	-4.3580	6.56E-06	3.28E-05
3S	329	208.6949	12.24355	9.8259	4.35E-23	2.18E-22
IR	238	316.8503	14.1002	-5.5921	1.12E-08	5.61E-08
3S5S	147	170.9809	11.10434	-2.1595	0.0154019	0.07700952

Supplementary table 4.4. Enrichment analysis by event type.

Enrichment and impoverishment of AS event types among genes with body lice specific AS events. Table shows all 5 AS types studied, exon skipping (ES), alternative 5' splice site (5S), alternative 3' splice site (3S), intron retention (IR) and alternative 3' and 5' splice sites (3S5S). Table shows expected and actual number of genes found for each event type. Last column shows the Bonferroni corrected p-value for each AS type.

AS event type	Body lice specific AS events per type	Expected number of body lice specific AS events from 10000 random samples	Standard Error	Z-score	One sided p-value	Bonferroni adjusted p-value
ES	742	540.3015	17.44294	11.56333	3.16E-31	1.58E-30
5S	488	568.337	17.55804	-4.57551	2.38E-06	1.19E-05
3S	420	321.8113	14.09827	6.964593	1.65E-12	8.23E-12
IR	313	489.227	16.78252	-10.50063	4.29E-26	2.15E-25
3S5S	220	263.3232	13.05875	-3.317561	0.000454036	0.00227018

Supplementary table 4.5. GO enrichment for body lice specific AS.

GO term categories enriched and impoverished among genes with body lice specific alternative splicing events. No terms were enriched or underrepresented in head lice. Table shows both expected and actual number of genes found for the categories. Last column shows the Benjamini-Hochberg corrected p-value for multiple testing.

Enriched GO Term	**GO Term**	**Genes with body lice specific AS events in GO Term**	**Expected no. of genes**	**Standard Error**	**Z-score**	**BH adjusted p-value**
GO:0007422	peripheral nervous system development	30	17.2124	3.2522	3.9319	0.0022
GO:0007435	salivary gland morphogenesis	26	15.2889	3.0898	3.4665	0.0069
GO:0007424	open tracheal system development	40	27.6003	4.0973	3.0262	0.0164
GO:0007391	dorsal closure	39	27.5647	4.1133	2.7800	0.0280
GO:0006355	regulation of transcription; DNA-dependent	50	37.1758	4.6985	2.7293	0.0280
GO:0030707	ovarian follicle cell development	25	16.807	3.2408	2.5280	0.0395
GO:0048102	autophagic cell death	24	16.0587	3.1591	2.5137	0.0395
Impoverished GO Term	**GO Term**	**Genes with body lice specific AS events in GO Term**	**Expected no. of genes**	**Standard Error**	**Z-score**	**BH adjusted p-value**
GO:0055114	oxidation-reduction process	26	41.73	5.0595	-3.10898	0.0164
GO:0006412	translation	12	20.8038	3.5644	-2.46986	0.0397

5 Alternative splicing: a potential source of functional innovation in the eukaryotic genome

5.1 Introduction

The first draft of the human genome sequence (Lander et al. 2001; Venter et al. 2001) was unveiled in February 2001 and surprisingly it was shown to contain ~23000 genes, only a fraction of the numbers of genes originally predicted (Crollius et al. 2000). To put this into perspective, there are ~20,000 genes in the genome of the nematode *C. elegans*. The lack of an association between gene number and organismal complexity has resulted in an increased interest in alternative splicing (AS) given it has been proposed to be a major factor in expanding the regulatory and functional complexity, protein diversity and organismal complexity of higher eukaryotes (Graveley 2001; Nilsen & Graveley 2010). However, despite the best efforts of many research groups we still understand very little about the actual role played by AS in the evolution of functional innovation -here understood as the appearance of novel functional transcripts- underpinning the increased organismal complexity observed.

Alternative splicing is a post-transcriptional process in eukaryotic organisms by which multiple distinct transcripts are produced from a single gene (Graveley 2001). Previous studies using high-throughput sequencing technology have reported that up to 92%~94% of human multi-exon genes undergo AS (Pan et al. 2008; Wang et al. 2008), often in a tissue/developmental stage-specific manner (Stamm et al. 2005; Wang et al. 2008). With the development and constant improvement of whole genome transcription profiling and bioinformatics algorithms, the ubiquity of AS in the mammalian genome began to become clear. The concept of one gene one protein gave way as evidence mounted for the high percentage of AS incidence in non-human species (Pan et al. 2008; Wang et al. 2008), such as mouse fruit fly (Graveley

et al. 2011), *Arabidopsis* (Filichkin et al. 2010) and other eukaryotes (N. Kim et al. 2007). Despite the advances in our understanding and characterisation of AS several questions remain unanswered. First, the large difference in transcript coverage between species has hampered direct comparisons of the prevalence of alternative splicing in different species (Nilsen & Graveley 2010). Secondly, even if comparable AS estimates between species could be obtained; it is unclear to what extent any changes in AS prevalence along evolution has contributed to overall protein diversity or rather reflect splicing noise. Finally, we understand very little about how AS has evolved through time and how this is related to functional parameters of genes. Here we review how alternative is regulated and recent progress in our understanding of the evolution of alternative splicing.

5.2 Alternative splicing and its regulation

In 1977, Chow et al. (Berget et al. 1977; Chow et al. 1977; Alt et al. 1980; Early et al. 1980) reported that 5' and 3' terminal sequences of several adenovirus 2 (Ad2) mRNAs varied, implying a new mechanism that the diversity of splicing patterns and the variety of recombined sequences generated during the synthesis of late Ad2 mRNAs, following this study, alternative splicing was also found in the gene encoding thyroid hormone calcitonin in mammalian cells. Subsequent studies revealed that many other genes were also able to generate more than one transcript by cutting-out different sections from its coding regions (reviewed in (Graveley 2001; Artamonova & Gelfand 2007)).

Depending on the location of the exonic segments cut out or if introns are left in, splicing events can be classified into four basic types (Figure 5.1). These four major modes of splicing are: (1) Exon skipping (2) intron retention (3) alternative 5' splicing site (5'ss) and (4) alternative 3' splicing site (3'ss) (Ast 2004; Malko et al. 2006). In addition, mutually exclusive exons, alternative initiation and alternative polyadenylation provide two other mechanisms for generating various transcript isoforms. Moreover, different types of alternative splicing can occur in a

combinatorial manner and one exon may be subject to more than one AS modes, for example, 5'ss and 3'ss at the same time (Figure 5.1). Prevalence of each type of AS has been found to vary between different taxa. Several studies have shown that exon skipping is common in metazoan genomes (E. Kim et al. 2007) whereas intron retention is the most common types of AS among plants (Wang & Brendel 2006; McGuire et al. 2008) and fungi (Kim et al. 2008a; McGuire et al. 2008).

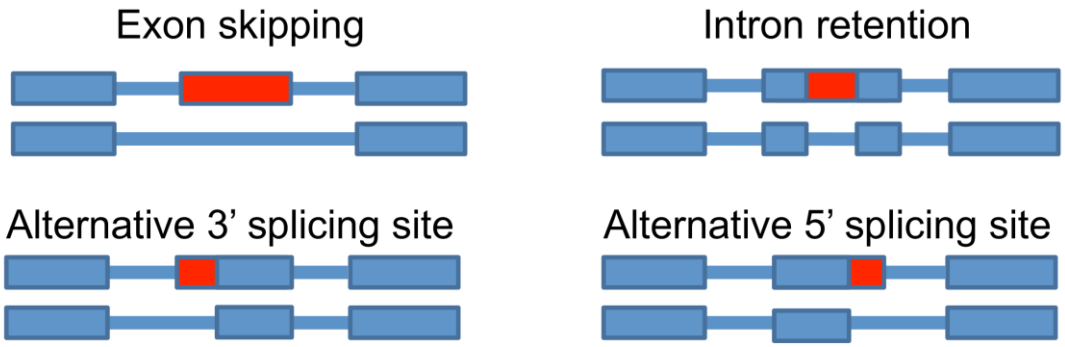


Figure 5.1. Different types of alternative splicing. The blue boxes are constitutive exons and alternatively spliced regions in red. Introns are represented by straight lines between boxes. Four types of common splicing events were identified: Exon skipping, intron retention, alternative 5' splicing site (5'ss) and alternative 3' splicing site (3'ss).

Alternative splicing is tightly regulated by *cis* elements as well as *trans*-acting factors that bind to these *cis* elements. *Trans*-acting factors, mainly RNA binding proteins, modulate the activity of the spliceosome and *cis* elements such as exonic splicing enhancers (ESE), exonic splicing silencers (ESS), intronic splicing enhancer (ISE) and intronic splicing silencers (ISS). Canonical mechanism of AS suggests that serine/arginine-rich (SR) proteins typically bind to ESEs, whereas heterogeneous nuclear ribonucleoproteins (hnRNP) tend to bind to ESSs or ISSs (Chen & Manley 2009). Given the crucial roles of these regulators in the splicing machinery, the *cis*- and *trans*-acting mutations, which disrupt the splicing code, are known to cause disease (reviewed in (Brinkman 2004; Venables 2006; Wang & Cooper 2007)). It has

been estimated that 15-60 % of mutations that cause in disease by affecting the splicing pattern of genes ((López-Bigas et al. 2005) and reviewed in (Wang & Cooper 2007)). Moreover, AS has also been shown to be regulated without the involvement of auxiliary splicing factors (Yu et al. 2008), AS may be also combined with other post-transcriptional events such as the use of multiple internal translation initiation sites, RNA editing, mRNA decay and microRNA binding and other non-coding RNAs (Hughes 2006; Luco & Misteli 2011), suggesting the existence of additional non-canonical mechanism of AS that are yet to be identified (Graveley 2009).

Recently, a direct role of histone modifications in alternative splicing has been reported, in which histone modification (H3-K27m3) affect the splicing outcome by influencing the recruitment of splicing regulators via a chromatin-binding protein in a number of human genes such as *FGFR2, TPM2, TPM1*, and *PKM2* (Luco et al. 2010). Moreover, it has been reported that CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing, providing the first evidence that the developmental regulation of splicing outcome through heritable epigenetic marks (Shukla et al. 2011). Additionally, non-coding RNAs also have emerged as key determinants of alternative splicing patterns (Luco & Misteli 2011). Therefore these findings reveal an additional epigenetic layer in the regulation of transcription and alternative splicing (Luco & Misteli 2011). Genome-wide genetic and epigenetic studies, therefore, have been proposed in at least 100 specific blood cell types (Adams et al. 2012), which will provide high quality reference epigenomes (using DNA methylation and histone marks assays) with detailed genetic and transcriptome data (whole genome sequencing, RNA-Seq and miRNA-Seq), providing us an opportunity to assess the genome-wide influence of epigenetic factors in the regulation of AS in specific blood cell types. We are expecting the rise of comparative epigenetics will provide different perspective of the evolution of transcriptome.

5.3 Identification of alternative splicing events

Alternative splicing is difficult to estimate from genomic parameters alone (Barash et al. 2010). A number of regulatory motifs for AS have been uncovered and but the presence of known alternative splicing motifs does not guarantee that a gene is actually alternatively spliced (Barash et al. 2010). Thus, alternative splicing patterns are generally assessed from examining transcript data. For any gene of interest, alternative splicing events can be identified by using reverse transcription polymerase chain reaction (RT-PCR) conducted on a complementary DNA (cDNA) library. Over the last decade, as high throughput transcriptome technologies have improved, it has become possible to assess alternative splicing patterns on a genome-wide scale. Three main sources of transcriptome data have been used to assess splicing patterns: expressed sequence tags (ESTs), splice-junction microarrays, and RNA sequencing (RNA-Seq).

The first wave of genome-wide transcriptome analysis consisted in direct sequencing cDNA and ESTs carried out at large scale (Sayers et al. 2009), which allowed alternative splicing events to be identified by aligning cDNA/EST sequences to the reference genome. ESTs are 200–800 nucleotide bases in length, unedited, randomly selected single-pass sequence reads derived from cDNA libraries (Nagaraj et al. 2007). Currently, there are eight million ESTs for human, including about one million sequences from cancer tissues, and about 71 million ESTs for around 2000 species in dbEST (Boguski et al. 1993). However, ESTs are based on low-throughput Sanger sequencing and are aggregated over a wide range of tissues, developmental states and diseases using widely different levels of sensitivity.

More recently, splice-junction microarrays and RNA-Seq have been increasingly used to quantitatively analyse alternative splicing events. Splicing microarrays target specific exons or exon-exon junctions with oligonucleotide probes. The fluorescent intensities of individual probes reflect the relative usage of alternatively splicing exons in different tissues and cell lines (Johnson et al. 2003). High-density splice-junction microarrays are a cost-effective way to assay previously

known exons and AS events with low false positive rate. The disadvantage is that it requires prior knowledge of existing AS variants and gene structures. More importantly unlike RNA-Seq and EST, microarrays do not provide additional sequence information.

RNA-Seq has emerged as a powerful technology for transcriptome analysis due to its ability to produce millions of short sequence reads (Wang et al. 2009; Robertson et al. 2010; Martin & Wang 2011). RNA-Seq experiments provide in-depth information on the transcriptional landscape (Wang et al. 2009). The ever increasing accumulation of high throughput data will continue to provide ever richer opportunities to investigate further aspects of AS such as low-frequency AS events as well as tissue-specific and/or development-specific AS events (Pan et al. 2008; Wang et al. 2008; Hawkins et al. 2010; Martin & Wang 2011; Ozsolak & Milos 2011). Earlier datasets consist of RNA read sequences of 50bp or less, limiting the information about combinations of AS events in a single transcript but it is likely that the length of short reads will continue to increase over the next decade. With the increasing capacity of next generation sequencing (RNA-Seq) the study of alternative spicing is likely to undergo a revolution (Mortazavi et al. 2008). The higher depth of sequencing of transcriptomes in human and other species has increased our understanding of the occurrence of AS event and AS expression patterns in different tissues (Wang et al. 2008; Kang et al. 2011), developmental stages (Graveley et al. 2011).

Transcript assembly of sequence-based technologies, such as ESTs and RNA-Seq, can use either align-then-assemble or assemble-then-align, depending on the quality of reference genome and sequence data (Martin & Wang 2011). An algorithm can be employed to detect AS event by comparing different transcripts. However, detecting AS isoforms, as opposed to single AS event, is still challenging because short sequences provide little information in terms of the combination of exons. Several applications have been developed for transcript assembly and AS isoform detection, different strategies and comparison of these applications have been reviewed previously (Martin & Wang 2011).

5.4 Prevalence of alternative splicing across eukaryotic genomes

Initial whole genome analyses suggested that 5%-30% of human genes were alternatively spliced (reviewed in (Artamonova & Gelfand 2007; Nilsen & Graveley 2010)). EST based AS databases identify AS events in 40-60% of human genes (N. Kim et al. 2007; Lee et al. 2007; Bhasi et al. 2009), however, recently this number has been revised over and over with the latest estimates showing that up to 94% of human multi-exon genes produce more than one transcript through alternative splicing (Pan et al. 2008; Wang et al. 2008). Understanding of how alternative splicing has changed over time could provide insights as to how alternative splicing has impacted on transcript and protein diversity and phenotype evolution (Nilsen & Graveley 2010). In fungi, AS is thought to be rare due to the low number of exons in yeast (Ast 2004). In plants it has been estimated that around 20% of genes undergo AS based on EST data (Wang & Brendel 2006), a recent study using RNA-Seq, however, suggests that at least approximately 42% of intron-containing genes in Arabidopsis are alternatively spliced (Filichkin et al. 2010). We are expecting significantly higher percentages of AS occurrence will be discovered from various eukaryotes given the in-depth studies of transcriptome using next-generation sequencing such as RNA-Seq are ongoing. A few studies have attempted to compare AS prevalence among different taxa with animals generally reported to have higher AS incidence than plants (Artamonova & Gelfand 2007) and vertebrates having a higher AS incidence than invertebrates (E. Kim et al. 2007). However, these studies are either based on limited data or failed to correct for differences in transcript coverage (Nilsen & Graveley 2010).

There are a number of databases that provide AS data for multiple species (N. Kim et al. 2007; Lee et al. 2007; Koscielny et al. 2009). However, these existing resources are primarily focused on animal species and have poor coverage for protist, fungal and plant genomes thus making it difficult to compare divergent taxa. Most importantly, none of these resources take into account the well-documented effects

of differential transcript coverage across genes within and between species which greatly influences AS detection rates (Brett et al. 2002; Kan et al. 2002; E. Kim et al. 2007; Nilsen & Graveley 2010). Random sampling has been used (E. Kim et al. 2007) and shown to minimize the bias of transcript coverage (Figure 5.2). We expect that similar strategies will be employed in future comparative AS data resources

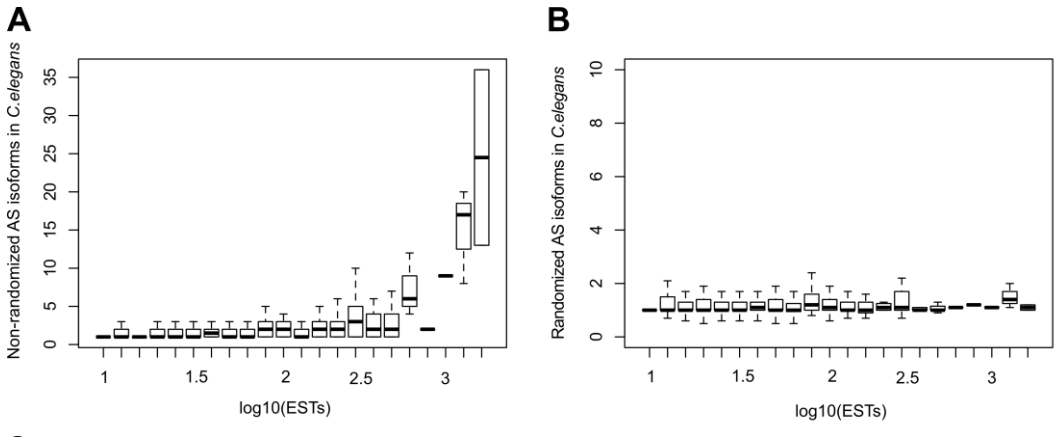


Figure 5.2. Total transcript number influences AS detection but bias can be corrected by using a sampling method. AS detection in genes divided by transcript coverage for the nematode (A and B) using the full transcript dataset (A) or a random sampling method (B).

5.5 Is alternative splicing functional or mostly just noise?

If an increase in AS levels in vertebrate species compared to invertebrates is confirmed, given the limitations of current proteomics resources, it is hard to assess the extent to which alternatively spliced transcripts are translated into an expanded proteome. The evolution of many phenotypes we most associate with human being such as longer lifespan, encephalization or even increased complexity have been accompanied by sharp reductions in effective population size, possibly explaining the proliferation of a variety of genomic features in more complex organisms ((Lynch & Conery 2003) but see (Whitney & Garland 2010)). Therefore, it is possible that

increased AS through evolution result from aberrant splicing and therefore it does not play any functional role (Xu 2003; Skotheim & Nees 2007; Kim et al. 2008b). If alternative splicing has increased along the phylogenetic tree and it is indeed functional, we can expect:

A) Transcripts should have a low incidence of premature stop codons which would render them vulnerable to nonsense mediated decay. Between 4%-35% of AS human transcripts have been found to contain a premature termination codon in human and mouse transcripts (Green et al. 2003; Lewis et al. 2003). These transcripts have been found to be enriched in non-conserved exons likely to cause frame shifts (Zhang et al. 2009). It is unknown whether the proportion of premature stop codon containing AS transcripts has changed along the phylogenetic tree.

B) It has been proposed that most low copy number alternative isoforms produced in human cells are likely to be non-functional (Su et al. 2006; Pickrell et al. 2010). A recent study has shown that although cancer-specific alternative splicing variants can be found, these events are mostly found as single copy events and thus unlikely to contribute to the core cancer transcriptome (Chen et al. 2011).

C) Conservation of alternative splicing events along evolution can be taken as an indicator of their functional role. Conservation levels of AS have been studied in many species. The estimation ranges from 11%-67% between human and mouse (Thanaraj et al. 2003; Pan et al. 2005; Mudge et al. 2011). Notably, major AS forms tend to be have higher conservation levels compared to minor forms. On the other hand, the conserved AS forms vary among different AS, for example, exon-skipping between *C. elegans* and *C. briggsae* has been shown more than 81% conservation level, compared to 28% for intron retention (Irimia et al. 2008; Irimia et al. 2009).

D) Presence of identifiable functional domains in AS areas may also be an indicator of functional relevance for AS transcripts (Chen et al. 2011). To our best knowledge there are no reports of the prevalence of functional domains in AS areas in model species. To examine the presence of functional domains in AS transcripts, we compiled a set of 267,996 AS events obtained from the analysis of 8,315,254 ESTs from normal human tissues. We found that about 50% of AS areas in human

contain known functional components using InterProScan (Zdobnov & Apweiler 2001) which contains 14 applications for the prediction of protein domains (Figure 5.3, see methods in (Chen et al. 2011)), suggesting a possible functional role for AS The extent of the variations in the prevalence of functional domains among AS areas between species remains to be explored but would provide additional insights on the evolution of AS.

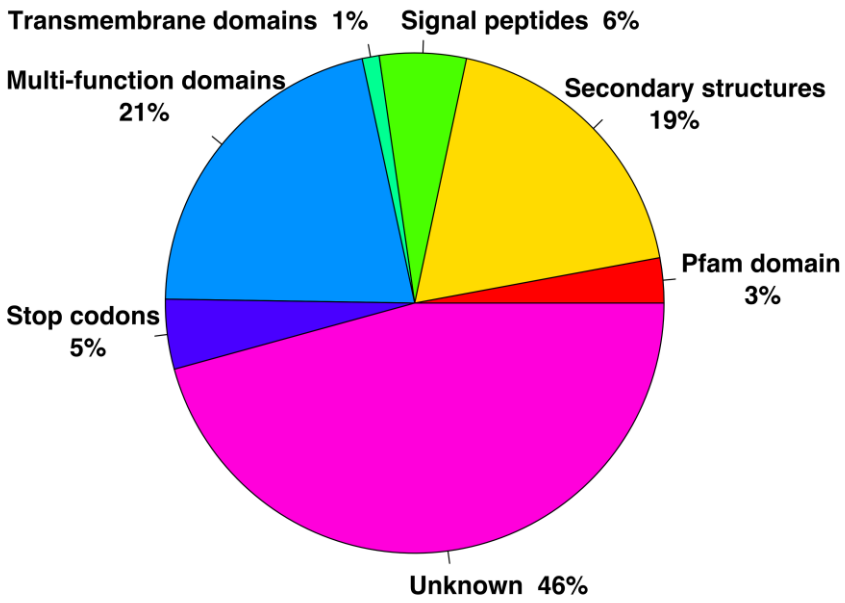


Figure 5.3. Percentage of AS areas contains identifiable functional domains, secondary structures and stop codons in human. Functional components were identified using InterProScan which contains 14 applications for the prediction of protein domains (Zdobnov & Apweiler 2001), including Pfam for the prediction of protein domains (Bateman et al. 2004), SignalP 3.0 for signal peptide predictions (Bendtsen et al. 2004) and TMHMM (Krogh et al. 2001) for the predictions of transmembrane domains. PSORT II (Nakai & Horton 1999) was used to identify the likely sub-cellular localization of protein products. Secondary protein structures were predicted by CLC Main Workbench 5.7, which is based on extracted protein sequences from the protein databank (http://www.rcsb.org/pdb/).

Taken together above observations suggest that although alternative splicing events are indeed conserved throughout evolution a significant proportion are not and some may result from noisy transcript splicing not contributing to the protein pool. However, until further studies using comparable AS indexes it will be impossible to estimate the extent to which increases in AS levels along the phylogenetic tree have impacted on the pool of functional transcripts.

5.6 Alternative splicing and gene duplication

Gene duplication (GD) is considered a prime source of functional innovation in the genome. Newly duplicated genes can evolve functional divergence (Long et al. 2003), and it is thought to be key in driving the evolution of developmental and morphological complexity in vertebrates (Dehal & Boore 2005). Alternative splicing, as a prevalent mechanism that also increases protein diversity has been proposed as a potential player in the evolution of eukaryotes (Graveley 2001; Nilsen & Graveley 2010). By examining the relationship between gene duplication and alternative splicing we can better understand the extent to which both mechanisms are equivalent means for protein diversification. Several studies have reported a negative correlation between AS and gene family size in human and mouse (Kopelman et al. 2005; Su et al. 2006; Jin et al. 2008; Nilsen & Graveley 2010) and worm (Hughes & Friedman 2008; Irimia et al. 2008) (Table 1). It is easy to lead to a conclusion that AS and GD are interchangeable and there is a universal negative correlation from worm to human. However, the relationship between the two variables is marginal at best and it is not consistent when including singleton genes which have a lower AS level compared to multi-gene families (Jin et al. 2008; Lin et al. 2008; Roux & Robinson-Rechavi 2011). Jin et al. (2008) suggested that singletons have more evolutionary constriction than duplicates which hampers their AS isoform gain. Consistent with this hypothesis, Lin et al. (2008) found that singletons differ from multi-gene families in several aspects suggesting that they have differing evolutionary paths. Even if we focus on multi-gene families only, a negative correlation between AS and gene family size may be explained or by-product of AS

and gene family size covariance with other factors. For example, gene age and biased duplication has been proposed to be the explanation (Roux & Robinson-Rechavi 2011). This study has cast doubt over the relationship between AS and GD and it may indeed provide support to the suggestion that AS and GD have little or no equivalence concerning effects on protein sequence, structure and function (Talavera et al. 2007). As most studies have examined a small number of model species it is difficult to generalise whether the extent of the link between AS and GD. In addition, the snapshot approach of comparing GFS and AS in a single genome might hide the true relationship between AS and GFS.

Table 1: Summary for the relationship between AS and GFS

Species	Data	Alternative splicing	Orthology	Bias control	Correlation	Reference
Human	Ensembl	ASD's AltSplice database	BLSATP	Exons, EST coverage, gene family size, isoform count	Negative correlation,	(Kopelman et al. 2005)
	NCBI, UCSC	GeneSplicer program	EnsMart	Remove garbage EST, EST coverage,	Negative correlation,	(Su et al. 2006)
	H-InvDB 5.0	H-InvDB 5.0	BLAST		Positive correlation when includes all gene families. Negative correlation within multi-gene families	(Jin et al. 2008)
Mouse	Ensembl	ASD's AltSplice database	BLSATP	Exons, EST coverage, gene family size, isoform count	Negative correlation,	(Kopelman et al. 2005)
	NCBI, UCSC	GeneSplicer program	EnsMart	Remove garbage EST, EST coverage,	Negative correlation,	(Su et al. 2006)
	Riken's FANTOM3	Riken's FANTOM3	BLAST		Positive correlation when includes all gene families. Negative correlation within multi-gene families	(Jin et al. 2008)
C.elegans	WormPep	WormPep	BLAST		Lower AS occurrence in multi-gene families	(Hughes & Friedman 2008)
Rice	TIGR 4.0	PASA program	BLASTP	Remove genes that lack transcript evidence	Multi-gene families have significantly higher AS incidence than singletons	(Alhashem et al. 2011)
Arabidopsis	TAIR7	TAIR7	TAIR7		Multi-gene families have significantly higher AS incidence than singletons	(Alhashem et al. 2011)

5.7 Alternative splicing's contribution to functional innovation

Alternative splicing has been hailed as the missing source of information in the genome accounting for the evolution of higher complexity despite as near static gene number in metazoans over the last 800 million years. Wegmann et al. (2008) found that width of gene expression is positively correlate to the number of new transcript isoforms, and proposed that the increase of gene expression breadth is essential for the gain new transcript isoforms that could be maintained by a new form of balancing selection. Moreover, experimental and bioinformatics analyses have shown that AS can generate a variety of functional mRNAs and protein products, displaying distinct stability properties, subcellular localization and function (Stamm et al. 2005) as well as in specific stages in cell differentiation (Heinzen et al. 2008), sex differentiation (Blekhman et al. 2010; Hartmann et al. 2011) and development (Stamm et al. 2005).

Single gene studies have provide examples where alternative splicing can lead to functional innovation before any events of gene duplication have taken place. One such example is that of Troponin I (TnI), which plays a key role in muscle contraction. In the vertebrate genome, TnI exists in three copies each expressed in a different muscle type (skeletal, fast and slow, and cardiac). In *Ciona*, one of the closest relatives of vertebrates TnI is present as a single gene. Interestingly, however, the *Ciona* gene produces three distinct alternatively spliced isoforms, each found to resemble the expression profile of one of the vertebrate genes suggesting that the specialisation of the TnI proteins to function in each muscle type preceded gene duplication events (MacLean et al. 1997). This pattern of alternative splice variants in ancestrally single genes resembling expression profiles of genes later duplicated has also been found in synapsin-2 genes in tetrapods (Yu et al. 2003) and *MITF* genes in teleost fish species (Lister et al. 2001; Altschmied et al. 2002). These examples suggest that alternative splicing can be a mechanism for functional

innovation preceding events of gene duplication through one of the three possible paths (Figure 5.4).

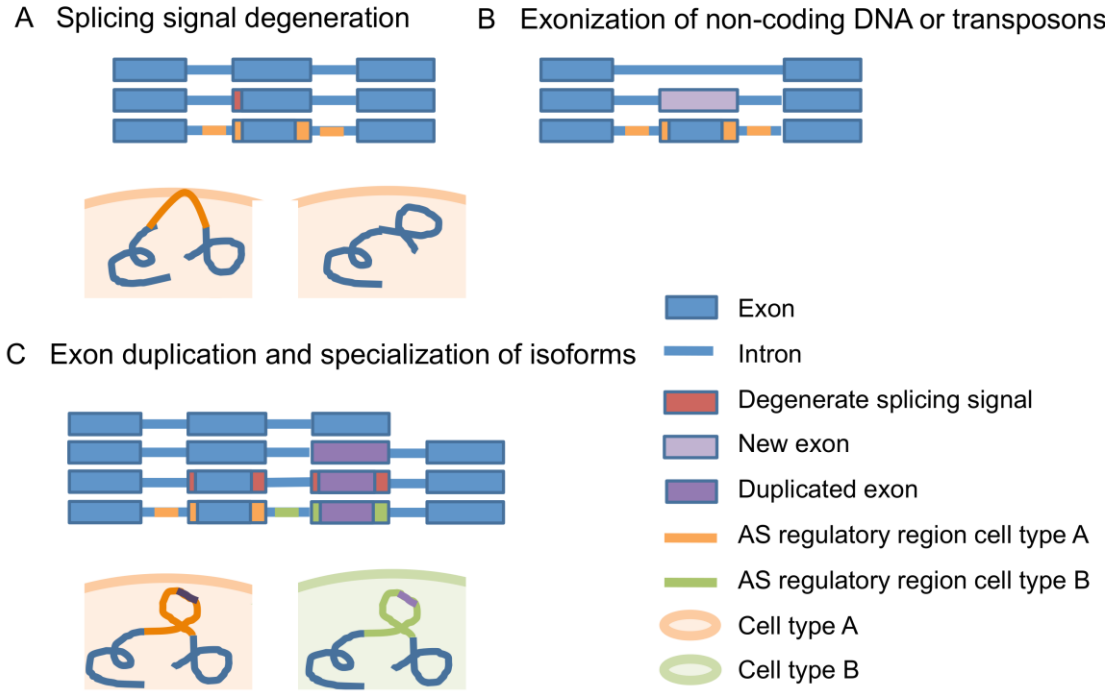


Figure 5.4. Novel AS variants can take on specialised or novel roles. Novel splicing variants can arise from A) mutations in the exon recognition site of a constitutive exon and subsequent acquisition of AS regulatory elements; B) Exonization of introns or intron regions or transposable elements with subsequent acquisition of AS regulatory regions. Novel proteins (AB) may interact with different proteins or localise in different sub-cellular regions. C) Exon duplication and subsequent specialization functional domains and AS regulatory regions. Resulting specialised proteins may take on partial roles relevant in different cell types or developmental stages or result in novel interactions and functions.

Genes may also further gain alternative splicing and regulation after duplication along with the complexity of the organ systems after the divergence of protochordates and vertebrates. Comparison between transcriptional factors *Pax* genes in vertebrates and amphioxus has shown that at least 52 reported alternative-splicing events in vertebrates compared to 23 events in amphioxus (Short & Holland 2008). Furthermore, vertebrate *Pax* genes have maintained most of their ancestral functions and also expanded their expression (Chen et al. 2010). Novel alternative splicing of *Pax* genes has been shown to modify the functional domain content (e.g.

DNA binding) and transactivation capacities of the resulting protein products (Short & Holland 2008). For example, a novel alternative transcript of *Pax3* can transactivate a cMET reporter construct in mouse (Barber et al. 1999). These additional isoforms of *Pax3* have been proposed to play a functional role in the acquisition of new roles at neural plate in vertebrates (Short & Holland 2008). Similarly, vertebrate-specific AS events of exon 5a in *Pax4* and *Pax6* have been linked to functional roles in the development of vertebrate eye (Singh et al. 2002; Short & Holland 2008). Therefore, it is reasonable to propose the hypothesis that, beside gene duplication, alternative splicing plays important roles in acquiring novel functions contributing to the complexity of the organ systems after the divergence of protochordates and vertebrates (Holland & Short 2010). The potential roles of the increasing prevalence of AS in vertebrates in functional innovation will be largely explored in more gene families or genome-wide level in the future, which will further our understanding of how AS contributes to functional innovation.

5.8 Conclusion

Here we have reviewed evidence from genome wide studies as well as possible avenues for future comparatives studies for the potential of alternative splicing as a source of functional innovation during the evolution of the eukaryotic genome. While it is now clear that AS is prevalent in the human genome, obstacles still remain in the assessment of how alternative splicing has evolved through time. The main obstacle lies in that while most other genomic features can be directly measured or estimated from genomic sequences alone, no accurate estimates of alternative splicing can be obtained from genomic sequence analysis. The reliance in transcript sequences availability to measure AS together with the strong bias brought by unequal transcript coverage has hampered the genome wide assessment of AS in all but a few model species and makes difficult any direct comparison between species. This has slowed down the study of how alternative splicing has evolved over time, how AS is regulated, and how it may relate to other genomic features and most crucially to phenotype. The ever increasing transcript profiling for many more

species combined with the use of comparable index estimates will allow addressing a number of evolutionary questions regarding the evolution of AS and its implications for the evolution of transcript diversity and functional innovation.

6 General discussion

Even decades after the first documented observation of AS in eukaryote cells, we haven't been able to answer how prevalent AS is in the highly diverse group of eukaryote species. As well, even though AS has been suggested as a possible source for functional diversification and innovation, we don't fully understand the relationship between AS, functional protein diversification and its effect on phenotypic innovations.

Alternative splicing event detection relies primarily in the analysis of transcript data and is highly dependent on transcript coverage. This is because as the number transcripts available for a particular gene increases, the proportion of AS events detected will also increase. Although it has long been speculated that alternative splicing is higher in multicellular eukaryotes compared to unicellular organisms, the lack of comparative estimates has hampered any efforts to confirm it (Nilsen & Graveley 2010).

The lack of comparable AS estimates and the underrepresentation of protist and fungi species in AS studies have also prevented efforts to investigate the evolution of alternative splicing in eukaryotic species (Nilsen & Graveley 2010; Schad et al. 2011). Furthermore, even in species where a high prevalence of AS has been confirmed, there is still intense debate about the roles AS and its evolution have played in the expansion of the functional proteome and the development and maintenance of complex phenotypes.

During my research I have addressed all these problems from different perspectives, studying the prevalence of AS in some of the taxonomic groups where alternative splicing has been most poorly addressed. I examined AS prevalence and levels in over a dozen protist and fungi species providing the most comprehensive study of alternative splicing in these taxonomic groups and the first comparative estimates correcting the distorting effects of differential transcript coverage. In addition, by conducting a detailed characterisation of AS in human head and body

lice, my results provide insights into the role AS may be playing in phenotypic variation and then engaging current discussions about the evolution of AS, its effect on protein diversification and the relationship between AS and complex phenotypes in general.

6.1 Characterising alternative splicing prevalence in protist species correcting for distorting effects of differential transcript coverage

Due to transcript coverage bias, many studies of alternative splicing prevalence based on transcript data presented conflicting results for same species (Brett et al. 2002; Kim et al. 2004). It was later established those results were strongly dependent on the transcript coverage of the libraries used (E. Kim et al. 2007). A random sampling normalisation method then proposed as a practical solution to the transcript coverage bias (E. Kim et al. 2007).

Here I have developed and applied a transcript coverage normalisation method to study the prevalence of AS in Protists. Protists represent a large and diverse set of organisms and have wide variations at the genomic level with some protists having extraordinary big genomes (Hackett et al. 2004; Bachvaroff & Place 2008), a propensity to have gene transfer (Bachvaroff & Place 2008; Lowe et al. 2011) and some of their genomes are known to contain several thousand introns (Gardner et al. 2002; Lu et al. 2007). Still very few characterisations of AS prevalence on protists species are available (Xiong et al. 2012; Lowe et al. 2011; Muhia et al. 2003; Coyne et al. 2008).

The results I here present show not only previous reports are limited in the number of species they report but they are probably underestimating the prevalence of AS. The average comparative prevalence of AS in 16 protist species I here present, are much higher than any other prevalence reported so far for individual species (Xiong et al. 2012). Thus suggesting the scarcity of transcript data in protists

has been limiting our understanding about how pervasive is this phenomenon in protists.

These results also contrast previous misconceptions related to low prevalence of AS in protists. At least two of the species in this study show levels of AS which are above of the levels calculated for *C. elegans* using the same method to correct for transcript coverage.

6.2 Alternative splicing prevalence in fungal species

Fungal models have played a crucial role in uncovering the organisation and function of eukaryotic genes and genomes. The analyses of fungal species allowed for the proposal of a link between specific proteins to individual genes which encoded them leading to the notion of "one gene one protein" (Beadle & Tatum 1941). When alternative splicing was first discovered, it was thought to be an exception rather than a rule in eukaryotic genomes. Later studies in human and other metazoan species changed this view revealing that alternative splicing is widespread in these genomes (Pan et al. 2008; Wang et al. 2008; N. Kim et al. 2007). Initial exploration in *S. cerevisiae*, the top fungi model species provided little evidence of alternative splicing with few early examples of AS in *S. cerevisiae* (three genes in yeast compared to 35% of genes in human, (reviewed in (Graveley 2001)). Moreover, characteristics of the cerevisiae genome, low intron prevalence (Kupfer et al. 2004), short introns (Long et al. 1997) and a simple splicing machinery (Clark et al. 2002), all justified poor prevalence of AS. Together, these observations in *S. cerevisiae* provided grounds to assume AS was a rare phenomenon not only in fungi but also in most of unicellular organisms. But the case of *S. cerevisiae* exemplifies the risk of generalising results from small sample of model species to a much broader group of organisms, as it is now the consensus view that the simple alternative splicing machinery is a derived characteristic, possibly resulting from relaxed constraints after an event of whole genome duplication in the *S. cerevisiae* lineage rather than an ancestral state.

Different studies in the last decade (Hirschman et al. 2006; Loftus et al. 2005; Marshall et al. 2013; McGuire et al. 2008; Zhao et al. 2013; Chang et al. 2010; Chang & Muddiman 2011) (also reviewed in (Kempken 2013; Kim et al. 2008a)), provided evidence supporting the idea of AS being a pervasive phenomenon in fungi. However, because they don't compensate for transcript coverage their results should not be used to compare prevalence of AS (E. Kim et al. 2007). In chapter 3, I present results of analyses on 23 fungi species, so far the study with the largest number of fungi species and also the first to provide information about the prevalence of AS using comparable indexes. My results confirm AS is likely to be a universal phenomenon in fungi genomes, with AS events detected in all the species. The results also confirm ES is uncommon, but together with the existence of 3S events, this study provides new evidence about how generalised is the use of exon definition in AS events of fungi species.

My analyses of AS in fungi taxa showed that the average prevalence of alternative splicing among the set of fungi species analysed is close to 20%, higher than previous estimates. Importantly, prevalence of AS was shown to be highly variable among 19 fungal species tested. Our studies, however, shed little light as to the factors explaining these variations. Phylogenetic relatedness was not shown as a relevant factor. This is perhaps not surprising as species analysed have in some cases diverged over 900 millions of years ago (Hedges et al. 2004).

Comparable prevalence interestingly shows some fungi species have similar or higher prevalence of AS to other eukaryotes (*C. elegans* and *D. melanogaster*). This challenges previous conceptions about the prevalence of AS in fungi species.

Even though several approaches were used, little evidence was found between AS and complex phenotypes tested. Only in the case of *U. maydis* it was possible to observe different prevalence of AS between transcript libraries associated with specific phenotypes. Nonetheless, functional characterisation showed alternatively spliced genes are not randomly distributed, in the case of *N. crassa* it was possible to identify a relationship between AS and genes associated with

translation. Notably, this GO category was also associated with alternatively spliced genes in protist species.

6.3 Alternative splicing in human head and body lice

A pervasive question about AS is how it is related with complex phenotypes. In chapter 4, I address this question through the analysis of transcription data from human head and body lice. Head and body lice offer a singular opportunity to study how differentiated AS patterns affect displayed phenotypes. Both types of lice have very similar transcriptional profiles (Olds et al. 2012). My work focused on assessing if differentiated patterns of AS could be associated with phenotypical differences in head and body lice. Alternative splicing has been previously associated with phenotypic plasticity in eusocial insects (Bonasio et al. 2012; Lyko et al. 2010) and has also been associated with insect organism development (reviewed in (Venables et al. 2012)), sex determination (reviewed in (Salz 2011; Venables et al. 2012)), transcription regulation (reviewed in (Venables et al. 2012)), with neural and immune system processes ((Watson et al. 2005) and reviewed in (Kurtz & Armitage 2006; Venables et al. 2012)).

My results revealed significant differences in the patterns of alternative splicing in head and body lice. Notably, although AS events specific to head and body lice were identified, the events specific to head lice more closely resembled alternative splicing patterns in the wider gene pool compared to AS events specific to body lice. Furthermore, while no single functional category was found to be enriched among genes with AS events specific to head lice, several functional categories all related with development and neural connectivity were found to be significantly associated with genes presenting AS evens specific to body lice. Although we recognise there are limitations associated with this kind of study. Our results suggest further studies will show AS has a close relationship with the morphological and ecological differences observed between head and body lice.

6.4 Alternative splicing and functional innovation in the eukaryote genome

Results presented chapter 4 reveal the potential role of alternative splicing in the evolution of adaptations of human lice in colonising a novel ecological niche after use of clothing became widespread among human populations. As there is no evidence as yet to support that the distinct phenotypes are associated with differences in gene content and that transcriptome analyses failed to find any evidence of differentially expressed genes, my results in principle raise the possibility that perhaps changes in alternative splicing may underlie the evolution of phenotypic differences in eukaryotic genomes sometimes preceding events of gene duplication and expression pattern differentiation. In chapter 5, I explore the potential role of alternative splicing in the evolution of novel functions in eukaryotic genomes. Although alternative splicing has been widely claimed to allow genomes to dramatically boost proteome size the consensus view remains that it is gene duplication, followed by subfunctionalisation and neofunctionalisation, the main driver of functional innovation in the genome.

Over the last 5 years a number of studies have specifically assessed the relationship between alternative splicing and gene duplication events. Most of these studies found an inverse relationship between alternative splicing and gene family size suggesting that, to some extent, these two processes are interchangeable (Su & Gu 2012) However this relationship remains controversial (Roux & Robinson-Rechavi 2011).

6.5 General conclusion

In this thesis I have characterised alternative splicing in a number of eukaryotic species of various taxa. Most of my efforts have been concentrated in characterising alternative splicing patterns in species where this process has remained understudied. In several protist and fungi species analysed in chapter 2-3

and section 2 of the appendices, my results constitute the first report of the assessment of alternative splicing. These analyses showed that alternative splicing is a ubiquitous feature in both of these taxa and that occurs at higher frequencies than previous reports. The high degree of variability in alternative splicing among species remains largely unexplained as phylogenetic relatedness does not account for the variations among species either in protist or fungi species. We also failed to recover any relationship between alternative splicing and some complex phenotypes. It is possible that as transcript data becomes available for a larger number of species, associations between this process and alternative splicing can be uncovered. The assessment of alternative splicing in human lice is also the first effort to characterise alternative splicing on a genome wide scale in this species and provides the first insights into the possible molecular mechanisms underlying the expression of distinct phenotypes allowing this organism to survive in two environments in its human host.

Important questions regarding alternative splicing remain unanswered. First what the contribution of alternative splicing to proteome size remains unclear as transcript coverage is still patchy in many taxa. Even in human where sequencing efforts have been more extensive, not all tissues or developmental stages have been assessed. Thus although estimates suggest that about 95% of human genes are alternatively spliced, it is unknown exactly how many distinct transcripts are in fact produced during the lifetime of a healthy being. Comparative estimates as the one used in this thesis, can help to draw estimates of alternative splicing across a whole genome. However, these estimates based on random samples of a set of transcripts per gene will invariably result in an underestimation of actual AS prevalence or in the number of alternative splicing isoforms produced.

Also, as several studies have shown, the presence of alternative spliced transcripts does not necessarily reflect the actual pool of functional proteins constituting the proteome. Several studies have reported that a significant proportion of alternative splicing events are in fact likely to result non-functional transcripts which are degraded before being translated (Xu 2003; Skotheim & Nees 2007; Kim et al. 2008b). This noise is particularly relevant in disease states where the presence of novel transcripts has sometimes been interpreted as evidence of a role of

alternative splicing in causing or maintaining the disease, rather than represent a marker of dis-regulation of transcription and RNA processing. Section 1 in the appendices examines this issue in cancer tissues finding that most transcripts which are cancer derived are in fact unique transcripts with increased levels of stop codons. Comparative studies using sets of orthologous genes in relatively closely related species have begun to shed light into the conservation rates of alternative splicing events, further supporting that a significant proportion of alternative splicing events are likely to be functional, even including a proportion of events which lead to the production of transcripts with premature stop codons (Ni et al. 2007). The use of comparative approaches to larger timespans and the analysis of the functional content (protein domain content and other functional features) in alternatively spliced regions along evolution might further shed light as to whether changes in the prevalence of alternative splicing along evolution results from noisier transcription and RNA processing or instead is likely to contribute to the diversification of protein products.

Third, what role alternative splicing patterns play in the sequence evolution of genes is not fully understood. Patterns of sequence evolution are fundamental to many areas of evolutionary biology. Understanding the determinants of variations in rates of sequence evolution is key to the notion of the molecular clock and its use in molecular phylogenetics which allow the resolution of topology of trees relating species to one another or the timing of radiations (dos Reis et al. 2012). It is now known that the presence of introns impacts the rates of sequence evolution, as splice enhancers result in higher synonymous site conservation than in other regions of the gene. Furthermore, alternative splicing creates disparities in the level of translation of certain segments of the coding region compared to constitutive exons. This has been speculated to result in reduced selective pressures on the sections not constitutively included in the mature transcripts (Iida & Akashi 2000). Indeed, results presented in sections 2 and 3 of the appendices suggest that alternative splicing might shape gene evolution both in terms of rates of nucleotide substitutions and potentially in polymorphic coding sequence deletion events. Future large scale analyses are likely to further unveil the extent to which alternative splicing is likely to shape the evolution of eukaryotic genes.

7 References

Adams, D., Altucci, L., Antonarakis, S.E., Ballesteros, J., Beck, S., Bird, A., Bock, C., Boehm, B., Campo, E., Caricasole, A., Dahl, F., Dermitzakis, E.T., Enver, T., Esteller, M., Estivill, X., FergusonSmith, A., Fitzgibbon, J., Flicek, P., Giehl, C., et al., 2012. BLUEPRINT to decode the epigenetic signature written in blood. *Nature biotechnology*, 30(3), pp.224–226.

Adams, M.D., 2000. The Genome Sequence of Drosophila melanogaster. *Science*, 287(5461), pp.2185–2195.

Alhashem, Y.N., Vinjamur, D.S., Basu, M., Klingmuller, U., Gaensler, K.M. & Lloyd, J.A., 2011. Transcription factors KLF1 and KLF2 positively regulate embryonic and fetal beta-globin genes through direct promoter binding. *J Biol Chem*, 286(28), pp.24819–24827.

Alt, F.W., Bothwell, A.L.M., Knapp, M., Siden, E., Mather, E., Koshland, M. & Baltimore, D., 1980. Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends. *Cell*, 20(2), pp.293–301.

Altschmied, J., Delfgaauw, J., Wilde, B., Duschl, J., Bouneau, L., Volff, J.N. & Schartl, M., 2002. Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish. *Genetics*, 161(1), pp.259–267.

Anczuków, O., Rosenberg, A.Z., Akerman, M., Das, S., Zhan, L., Karni, R., Muthuswamy, S.K. & Krainer, A.R., 2012. The splicing factor SRSF1 regulates apoptosis and proliferation to promote mammary epithelial cell transformation. *Nature structural & molecular biology*, 19(2), pp.220–8.

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., Brzezinski, M.A., Chaal, B.K., Chiovitti, A., Davis, A.K., Demarest, M.S., Detter, J.C., Glavina, T., Goodstein, D., Hadi, M.Z., et al., 2004. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science (New York, N.Y.)*, 306(5693), pp.79–86.

Artamonova, I.I. & Gelfand, M.S., 2007. Comparative genomics and evolution of alternative splicing: the pessimists' science. *Chemical reviews*, 107(8), pp.3407–30.

Ast, G., 2004. How did alternative splicing evolve? *Nature reviews. Genetics*, 5(10), pp.773–82.

Aurrecoechea, C., Brestelli, J., Brunk, B.P., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O.S., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J.C., Kraemer, E.T., Li, W., Miller, J.A., Nayak, V., Pennington, C., Pinney, D.F., et al., 2010. EuPathDB: a portal to eukaryotic pathogen databases. *Nucleic acids research*, 38(Database issue), pp.D415–9.

Baba, Y., Shimonaka, A., Koga, J., Kubota, H. & Kono, T., 2005. Alternative splicing produces two endoglucanases with one or two carbohydrate-binding modules in Mucor circinelloides. *Journal of bacteriology*, 187(9), pp.3045–51.

Bachvaroff, T.R. & Place, A.R., 2008. From stop to start: tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate Amphidinium carterae. *PloS one*, 3(8), p.e2929.

Bacot, A., 2009. A Contribution to the Bionomics of Pediculus humanus (Vestimenti) and Pediculus capitis. *Parasitology*, 9(02), p.228.

Baehrecke, E.H., 2002. How death shapes life during development. *Nature reviews. Molecular cell biology*, 3(10), pp.779–87.

Baer, K., Klotz, C., Kappe, S.H.I., Schnieder, T. & Frevert, U., 2007. Release of hepatic Plasmodium yoelii merozoites into the pulmonary microvasculature. *PLoS pathogens*, 3(11), p.e171.

Baker, B.S. & Wolfner, M.F., 1988. A molecular analysis of doublesex, a bifunctional gene that controls both male and female sexual differentiation in Drosophila melanogaster. *Genes & Development*, 2(4), pp.477–489.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J. & Frey, B.J., 2010. Deciphering the splicing code. *Nature*, 465(7294), pp.53–9.

Barbazuk, W.B., Fu, Y. & McGinnis, K.M., 2008. Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome research*, 18(9), pp.1381–92.

Barber, T.D., Barber, M.C., Cloutier, T.E. & Friedman, T.B., 1999. PAX3 gene structure, alternative splicing and evolution. *Gene*, 237(2), pp.311–319.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., Kim, T., Misquitta-Ali, C.M., Wilson, M.D., Kim, P.M., Odom, D.T., Frey, B.J. & Blencowe, B.J., 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science (New York, N.Y.)*, 338(6114), pp.1587–93.

Barchuk, A.R., Cristino, A.S., Kucharski, R., Costa, L.F., Simões, Z.L.P. & Maleszka, R., 2007. Molecular determinants of caste differentiation in the highly eusocial honeybee Apis mellifera. *BMC developmental biology*, 7, p.70.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats, C. & Eddy, S.R., 2004. The Pfam protein families database. *Nucleic Acids Research*, 32, pp.D138–D141.

Beadle, G. & Tatum, E., 1941. Genetic control of biochemical reactions in Neurospora. *Proc Natl Acad Sci U S A*, 27(11), pp.499–506.

Bendtsen, J.D., Nielsen, H., von Heijne, G. & Brunak, S., 2004. Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology*, 340(4), pp.783–795.

Berget, S.M., Moore, C., Sharp, P.A. & C., M., 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. 1977. *Reviews in medical virology*, 10(6), pp.356–362.

Bermingham, J.R. & Scott, M.P., 1988. Developmentally regulated alternative splicing of transcripts from the Drosophila homeotic gene Antennapedia can produce four different proteins. *The EMBO journal*, 7(10), pp.3211–22.

Bhasi, A., Philip, P., Sreedharan, V.T. & Senapathy, P., 2009. AspAlt: A tool for inter-database, inter-genomic and user-specific comparative analysis of alternative transcription and alternative splicing in 46 eukaryotes. *Genomics*, 94(1), pp.48–54.

Black, D.L., 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annual review of biochemistry*, 72, pp.291–336.

Black, D.L., 2000. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*, 103(3), pp.367–70.

Blekhman, R., Marioni, J.C., Zumbo, P., Stephens, M. & Gilad, Y., 2010. Sex-specific and lineage-specific alternative splicing in primates. *Genome research*, 20(2), pp.180–189.

Boettner, D.R., Huston, C.D., Linford, A.S., Buss, S.N., Houpt, E., Sherman, N.E. & Petri, W.A., 2008. Entamoeba histolytica phagocytosis of human erythrocytes involves PATMK, a member of the transmembrane kinase family. *PLoS pathogens*, 4(1), p.e8.

Boguski, M.S., Lowe, T.M. & Tolstoshev, C.M., 1993. dbEST--database for "expressed sequence tags". *Nature genetics*, 4(4), pp.332–3.

Bonasio, R., Li, Q., Lian, J., Mutti, N.S., Jin, L., Zhao, H., Zhang, P., Wen, P., Xiang, H., Ding, Y., Jin, Z., Shen, S.S., Wang, Z., Wang, W., Wang, J., Berger, S.L., Liebig, J., Zhang, G. & Reinberg, D., 2012. Genome-wide and caste-

specific DNA methylomes of the ants Camponotus floridanus and Harpegnathos saltator. *Current biology*, 22(19), pp.1755–64.

Bousema, T. & Drakeley, C., 2011. Epidemiology and infectivity of Plasmodium falciparum and Plasmodium vivax gametocytes in relation to malaria control and elimination. *Clinical microbiology reviews*, 24(2), pp.377–410.

Bowman, a S., Coons, L.B., Needham, G.R. & Sauer, J.R., 1997. Tick saliva: recent advances and implications for vector competence. *Medical and veterinary entomology*, 11(3), pp.277–85.

Bozzaro, S. & Eichinger, L., 2011. The professional phagocyte Dictyostelium discoideum as a model host for bacterial pathogens. *Current drug targets*, 12(7), pp.942–54.

Breitbart, R.E., Andreadis, A. & Nadal-Ginard, B., 1987. Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. *Annual review of biochemistry*, 56, pp.467–95.

Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbrück, S., Krueger, S., Reich, J. & Bork, P., 2000. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS letters*, 474(1), pp.83–6.

Brett, D., Pospisil, H., Valcárcel, J., Reich, J. & Bork, P., 2002. Alternative splicing and genome complexity. *Nature genetics*, 30(1), pp.29–30.

Brinkman, B.M.N., 2004. Splice variants as cancer biomarkers. *Clinical biochemistry*, 37(7), pp.584–594.

Brown, D.W., Butchko, R. a. E. & Proctor, R.H., 2008. Genomic analysis of Fusarium verticillioides. *Food Additives & Contaminants: Part A*, 25(9), pp.1158–1165.

Brucker, R.M. & Bordenstein, S.R., 2012. Speciation by symbiosis. *Trends in ecology & evolution*, 27(8), pp.443–51.

Busvine, J.R.J., 1948. The “head” and “body” races of Pediculus humanus L. *Parasitology*, 39(1-2), pp.1–16.

Buxton, P., 1947. *The Louse, an account of the lice which infest man, their medical importance and control*, London: Edward Arnold & Co.

Cacciò, S.M. & Ryan, U., 2008. Molecular epidemiology of giardiasis. *Molecular and biochemical parasitology*, 160(2), pp.75–80.

Caenorhabditis elegans Sequencing Consortium, 1998. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science*, 282(5396), pp.2012–2018.

Campion, C., Massiot, P. & Rouxel, F., 1997. Aggressiveness and production of cell-wall degrading enzymes by Pythium violae, Pythium sulcatum and Pythium ultimum, responsible for cavity spot on carrots. *European Journal of Plant Pathology*, pp.725–735.

Carlton, J.M., Adams, J.H., Silva, J.C., Bidwell, S.L., Lorenzi, H., Caler, E., Crabtree, J., Angiuoli, S. V, Merino, E.F., Amedeo, P., Cheng, Q., Coulson, R.M.R., Crabb, B.S., Del Portillo, H.A., Essien, K., Feldblyum, T. V, Fernandez-Becerra, C., Gilson, P.R., Gueye, A.H., et al., 2008. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. *Nature*, 455(7214), pp.757–63.

Carlton, J.M., Angiuoli, S. V, Suh, B.B., Kooij, T.W., Pertea, M., Silva, J.C., Ermolaeva, M.D., Allen, J.E., Selengut, J.D., Koo, H.L., Peterson, J.D., Pop, M., Kosack, D.S., Shumway, M.F., Bidwell, S.L., Shallom, S.J., van Aken, S.E., Riedmuller, S.B., Feldblyum, T. V, et al., 2002. Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. *Nature*, 419(6906), pp.512–9.

Caron, D.A., Worden, A.Z., Countway, P.D., Demir, E. & Heidelberg, K.B., 2009. Protists are microbes too: a perspective. *The ISME journal*, 3(1), pp.4–12.

Chang, K.-Y., Georgianna, D.R., Heber, S., Payne, G.A. & Muddiman, D.C., 2010. Detection of alternative splice variants at the proteome level in Aspergillus flavus. *Journal of proteome research*, 9(3), pp.1209–17.

Chang, K.-Y. & Muddiman, D.C., 2011. Identification of alternative splice variants in Aspergillus flavus through comparison of multiple tandem MS search algorithms. *BMC genomics*, 12(1), p.358.

Chang, Y.-C., Chang, T.-J., Jiang, Y.-D., Kuo, S.-S., Lee, K.-C., Chiu, K.C. & Chuang, L.-M., 2007. Association study of the genetic polymorphisms of the transcription factor 7-like 2 (TCF7L2) gene and type 2 diabetes in the Chinese population. *Diabetes*, 56(10), pp.2631–7.

Chen, L., Tovar-Corona, J.M. & Urrutia, A.O., 2012. Alternative splicing: a potential source of functional innovation in the eukaryotic genome. *International journal of evolutionary biology*, 2012, p.596274.

Chen, L., Tovar-Corona, J.M. & Urrutia, A.O., 2011. Increased levels of noisy splicing in cancers, but not for oncogene-derived transcripts. *Human molecular genetics*, 20(22), pp.4422–4429.

Chen, L., Zhang, Q., Wang, W. & Wang, Y., 2010. Spatiotemporal expression of Pax genes in amphioxus: insights into Pax-related organogenesis and evolution. *Science China. Life sciences*, 53(8), pp.1031–1040.

Chen, M. & Manley, J.L., 2009. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature reviews. Molecular cell biology*, 10(11), pp.741–54.

Chow, L.T., Gelinas, R.E., Broker, T.R. & Roberts, R.J., 1977. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell*, 12(September), pp.1–8.

Clark, T.A., Sugnet, C.W. & Ares, M., 2002. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science (New York, N.Y.)*, 296(5569), pp.907–10.

Cogan, J., 1997. A novel mechanism of aberrant pre-mRNA splicing in humans. *Human Molecular Genetics*, 6(6), pp.909–912.

Corliss, J., 2002. Biodiversity and biocomplexity of the protists and an overview of their significant roles in maintenance of our biosphere. *Acta Protozoologica*, 41(3), pp.199–219.

Coyne, R.S., Thiagarajan, M., Jones, K.M., Wortman, J.R., Tallon, L.J., Haas, B.J., Cassidy-Hanley, D.M., Wiley, E. a, Smith, J.J., Collins, K., Lee, S.R., Couvillion, M.T., Liu, Y., Garg, J., Pearlman, R.E., Hamilton, E.P., Orias, E., Eisen, J. a & Methé, B. a, 2008. Refined annotation and assembly of the Tetrahymena thermophila genome sequence through EST analysis, comparative genomic hybridization, and targeted gap closure. *BMC genomics*, 9, p.562.

Crollius, H.R., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., Saurin, W. & Weissenbach, J., 2000. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nature Genetics*, 25(2), pp.235–238.

Cuperlovic-Culf, M., Belacel, N., Culf, A.S. & Ouellette, R.J., 2006. Microarray analysis of alternative splicing. *Omics : a journal of integrative biology*, 10(3), pp.344–57.

Dehal, P. & Boore, J.L., 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *Plos Biology*, 3(10), pp.1700–1708.

Deutsch, M. & Long, M., 1999. Intron-exon structures of eukaryotic model organisms. *Nucleic acids research*, 27(15), pp.3219–28.

Dionne, S.O., Podany, A.B., Ruiz, Y.W., Ampel, N.M., Galgiani, J.N. & Lake, D.F., 2006. Spherules derived from Coccidioides posadasii promote human dendritic cell maturation and activation. *Infection and immunity*, 74(4), pp.2415–22.

Dong, Y., Taylor, H.E. & Dimopoulos, G., 2006. AgDscam, a hypervariable immunoglobulin domain-containing receptor of the Anopheles gambiae innate immune system. *PLoS biology*, 4(7), p.e229.

Drali, R., Boutellis, A., Raoult, D., Rolain, J.M. & Brouqui, P., 2013. Distinguishing body lice from head lice by multiplex real-time PCR analysis of the Phum_PHUM540560 gene. *PloS one*, 8(2), p.e58088.

Dubey, J.P., 2004. Toxoplasmosis - a waterborne zoonosis. *Veterinary parasitology*, 126(1-2), pp.57–72.

DuPont, H.L., Chappell, C.L., Sterling, C.R., Okhuysen, P.C., Rose, J.B. & Jakubowski, W., 1995. The infectivity of Cryptosporidium parvum in healthy volunteers. *The New England journal of medicine*, 332(13), pp.855–9.

Durden, L.A. & Musser, G.G., 1994. The sucking lice (Insecta, Anoplura) of the world; a taxonomic checklist with records of mammalian hosts and geographical distributions. *Bulletin of the American Museum of Natural History*, (218), pp.1–90.

Duvick, J., Fu, A., Muppirala, U., Sabharwal, M., Wilkerson, M.D., Lawrence, C.J., Lushbough, C. & Brendel, V., 2008. PlantGDB: a resource for comparative plant genomics. *Nucleic acids research*, 36(Database issue), pp.D959–65.

Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R. & Hood, L., 1980. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell*, 20(2), pp.313–9.

Eizaguirre, C., Lenz, T.L., Traulsen, A. & Milinski, M., 2009. Speciation accelerated and stabilized by pleiotropic major histocompatibility complex immunogenes. *Ecology letters*, 12(1), pp.5–12.

Ermakova, E.O., Nurtdinov, R.N. & Gelfand, M.S., 2006. Fast rate of evolution in alternatively spliced coding regions of mammalian genes. *BMC genomics*, 7, p.84.

Escalante, R., Moreno, N. & Sastre, L., 2003. Dictyostelium discoideum developmentally regulated genes whose expression is dependent on MADS box transcription factor SrfA. *Eukaryotic cell*, 2(6), pp.1327–35.

Fields, C., Adams, M.D., White, O. & Venter, J.C., 1994. How many genes in the human genome? *Nature genetics*, 7(3), pp.345–6.

Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E., Wong, W.-K. & Mockler, T.C., 2010. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome research*, 20(1), pp.45–58.

Fischer-Parton, S., Parton, R.M., Hickey, P.C., Dijksterhuis, J., Atkinson, H. a & Read, N.D., 2000. Confocal microscopy of FM4-64 as a tool for analysing endocytosis and vesicle trafficking in living fungal hyphae. *Journal of microscopy*, 198(Pt 3), pp.246–59.

Florens, L., Washburn, M.P., Raine, J.D., Anthony, R.M., Grainger, M., Haynes, J.D., Moch, J.K., Muster, N., Sacci, J.B., Tabb, D.L., Witney, A.A., Wolters, D., Wu, Y., Gardner, M.J., Holder, A.A., Sinden, R.E., Yates, J.R. & Carucci, D.J., 2002. A proteomic view of the Plasmodium falciparum life cycle. *Nature*, 419(6906), pp.520–6.

Gabellini, D., D'Antona, G., Moggio, M., Prelle, A., Zecca, C., Adami, R., Angeletti, B., Ciscato, P., Pellegrino, M.A., Bottinelli, R., Green, M.R. & Tupler, R., 2006. Facioscapulohumeral muscular dystrophy in mice overexpressing FRG1. *Nature*, 439(7079), pp.973–7.

Galagan, J.E., Calvo, S.E., Borkovich, K. a, Selker, E.U., Read, N.D., Jaffe, D., FitzHugh, W., Ma, L.-J., Smirnov, S., Purcell, S., Rehman, B., Elkins, T., Engels, R., Wang, S., Nielsen, C.B., Butler, J., Endrizzi, M., Qui, D., Ianakiev, P., et al., 2003. The genome sequence of the filamentous fungus Neurospora crassa. *Nature*, 422(6934), pp.859–68.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., Paulsen, I.T., James, K., Eisen, J. a, Rutherford, K., Salzberg, S.L., Craig, A., Kyes, S., Chan, M.-S., Nene, V., et al., 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419(6906), pp.498–511.

De Geer, C., 1767. *Mémoires pour servir à l'histoire des insectes*, Stockholm: Pierre Hasselberg.

Gelfman, S., Burstein, D., Penn, O., Savchenko, A., Amit, M., Schwartz, S., Pupko, T. & Ast, G., 2012. Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. *Genome research*, 22(1), pp.35–50.

Gerstein, M.B., Lu, Z.J., Van Nostrand, E.L., Cheng, C., Arshinoff, B.I., Liu, T., Yip, K.Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., Alves, P., Chateigner, A., Perry, M., Morris, M., Auerbach, R.K., Feng, X., Leng, J., Vielle, A., Niu, W., et al., 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science (New York, N.Y.)*, 330(6012), pp.1775–87.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. & Oliver, S.G., 1996. Life with 6000 genes. *Science (New York, N.Y.)*, 274(5287), pp.546, 563–7.

Goodarzi, M.O. & Rotter, J.I., 2007. Testing the gene or testing a variant? The case of TCF7L2. *Diabetes*, 56(10), pp.2417–9.

Grant, S.F.A., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A., Styrkarsdottir, U., Magnusson, K.P., Walters, G.B., Palsdottir, E., Jonsdottir, T., Gudmundsdottir, T., Gylfason, A., Saemundsdottir, J., Wilensky, R.L., et al., 2006. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nature genetics*, 38(3), pp.320–3.

Graveley, B.R., 2001. Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics*, 17(2), pp.100–107.

Graveley, B.R., 2009. Alternative splicing: regulation without regulators. *Nature structural & molecular biology*, 16(1), pp.13–5.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W., Brown, J.B., Cherbas, L., Davis, C.A., Dobin, A., Li, R., Lin, W., Malone, J.H., Mattiuzzo, N.R., Miller, D., et al., 2011. The developmental transcriptome of Drosophila melanogaster. *Nature*, 471(7339), pp.473–479.

Gray, J., 1998. Review The ecology of ticks transmitting Lyme borreliosis. *Experimental & applied acarology*, 22, pp.249–258.

Green, R.E., Lewis, B.P., Hillman, R.T., Blanchette, M., Lareau, L.F., Garnett, A.T., Rio, D.C. & Brenner, S.E., 2003. Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. *Bioinformatics*, 19, pp.i118–i121.

Grigoriev, I. V, Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D., Kuo, A., Minovitsky, S., Nikitin, R., Ohm, R.A., Otillar, R., Poliakov, A., Ratnere, I., Riley, R., Smirnova, T., Rokhsar, D. & Dubchak, I., 2012. The genome portal of the Department of Energy Joint Genome Institute. *Nucleic acids research*, 40(Database issue), pp.D26–32.

Grisdale, C.J., Bowers, L.C., Didier, E.S. & Fast, N.M., 2013. Transcriptome analysis of the parasite Encephalitozoon cuniculi: an in-depth examination of pre-mRNA splicing in a reduced eukaryote. *BMC genomics*, 14(1), p.207.

Haber, D.A., Sohn, R.L., Buckler, A.J., Pelletier, J., Call, K.M. & Housman, D.E., 1991. Alternative splicing and genomic structure of the Wilms tumor gene WT1. *Proceedings of the National Academy of Sciences*, 88(21), pp.9618–9622.

Hackett, J.D., Anderson, D.M., Erdner, D.L. & Bhattacharya, D., 2004. Dinoflagellates: a remarkable evolutionary experiment. *American journal of botany*, 91(10), pp.1523–34.

Haerty, W., Jagadeeshan, S., Kulathinal, R.J., Wong, A., Ravi Ram, K., Sirot, L.K., Levesque, L., Artieri, C.G., Wolfner, M.F., Civetta, A. & Singh, R.S., 2007. Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila. *Genetics*, 177(3), pp.1321–35.

Hahn, M.W. & Wray, G. a, 2002. The g-value paradox. *Evolution & development*, 4(2), pp.73–5.

Hanke, J., Brett, D., Zastrow, I., Aydin, A., Delbrück, S., Lehmann, G., Luft, F., Reich, J. & Bork, P., 1999. Alternative splicing of human genes: more the rule than the exception? *Trends in genetics : TIG*, 15(10), pp.389–90.

Haque, R., Mondal, D., Kirkpatrick, B.D., Akther, S., Farr, B.M., Sack, R.B. & Petri, W. a, 2003. Epidemiologic and clinical characteristics of acute diarrhea with emphasis on Entamoeba histolytica infections in preschool children in an urban slum of Dhaka, Bangladesh. *The American journal of tropical medicine and hygiene*, 69(4), pp.398–405.

Harr, B. & Turner, L.M., 2010. Genome-wide analysis of alternative splicing evolution among Mus subspecies. *Molecular ecology*, 19 Suppl 1(s1), pp.228–39.

Hartmann, B., Castelo, R., Minana, B., Peden, E., Blanchette, M., Rio, D.C., Singh, R. & Valcarcel, J., 2011. Distinct regulatory programs establish widespread sex-specific alternative splicing in Drosophila melanogaster. *RNA*, 17(3), pp.453–468.

Hastings, G.A. & Emerson, C.P., 1991. Myosin functional domains encoded by alternative exons are expressed in specific thoracic muscles of Drosophila. *The Journal of cell biology*, 114(2), pp.263–76.

Hatton, A.R., Subramaniam, V. & Lopez, A.J., 1998. Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. *Molecular cell*, 2(6), pp.787–96.

Hattori, D., Demir, E., Kim, H.W., Viragh, E., Zipursky, S.L. & Dickson, B.J., 2007. Dscam diversity is essential for neuronal wiring and self-recognition. *Nature*, 449(7159), pp.223–7.

Hawkins, R.D., Hon, G.C. & Ren, B., 2010. Next-generation genomics: an integrative approach. *Nature reviews. Genetics*, 11(7), pp.476–486.

Hedges, S.B., Blair, J.E., Venturi, M.L. & Shoe, J.L., 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC evolutionary biology*, 4, p.2.

Hedges, S.B., Dudley, J. & Kumar, S., 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics (Oxford, England)*, 22(23), pp.2971–2.

Heinzen, E.L., Ge, D., Cronin, K.D., Maia, J.M., Shianna, K. V, Gabriel, W.N., Welsh-Bohmer, K.A., Hulette, C.M., Denny, T.N. & Goldstein, D.B., 2008. Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS biology*, 6(12), p.e1.

Helgason, A., Pálsson, S., Thorleifsson, G., Grant, S.F.A., Emilsson, V., Gunnarsdottir, S., Adeyemo, A., Chen, Y., Chen, G., Reynisdottir, I., Benediktsson, R., Hinney, A., Hansen, T., Andersen, G., Borch-Johnsen, K., Jorgensen, T., Schäfer, H., Faruque, M., Doumatey, A., et al., 2007. Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nature genetics*, 39(2), pp.218–25.

Hirschman, J.E., Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hong, E.L., Livstone, M.S., Nash, R., Park, J., Oughtred, R., Skrzypek, M., Starr, B., Theesfeld, C.L., Williams, J., Andrada, R., Binkley, G., Dong, Q., et al., 2006. Genome Snapshot: a new resource at the Saccharomyces Genome Database (SGD) presenting an overview of the Saccharomyces cerevisiae genome. *Nucleic acids research*, 34(Database issue), pp.D442–445.

Holland, L.Z. & Short, S., 2010. Alternative splicing in development and function of chordate endocrine systems: a focus on Pax genes. *Integrative and comparative biology*, 50(1), pp.22–34.

Holste, D. & Ohler, U., 2008. Strategies for identifying RNA splicing regulatory motifs and predicting alternative splicing events. *PLoS computational biology*, 4(1), p.e21.

House, A.E. & Lynch, K.W., 2008. Regulation of alternative splicing: more than just the ABCs. *The Journal of biological chemistry*, 283(3), pp.1217–21.

Hsu, S.-N., Yonekura, S., Ting, C.-Y., Robertson, H.M., Iwai, Y., Uemura, T., Lee, C.-H. & Chiba, A., 2009. Conserved alternative splicing and expression patterns of arthropod N-cadherin. *PLoS genetics*, 5(4), p.e1000441.

Hua, Y., Sahashi, K., Hung, G., Rigo, F., Passini, M. a, Bennett, C.F. & Krainer, A.R., 2010. Antisense correction of SMN2 splicing in the CNS rescues necrosis in a type III SMA mouse model. *Genes & development*, 24(15), pp.1634–44.

Hughes, A.L. & Friedman, R., 2008. Alternative splicing, gene duplication and connectivity in the genetic interaction network of the nematode worm Caenorhabditis elegans. *Genetica*, 134(2), pp.181–186.

Hughes, T.A., 2006. Regulation of gene expression by alternative untranslated regions. *Trends in genetics : TIG*, 22(3), pp.119–122.

Iida, K. & Akashi, H., 2000. A test of translational selection at "silent" sites in the human genome: base composition comparisons in alternatively spliced genes. *Gene*, 261(1), pp.93–105.

International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), pp.931–945.

Iriko, H., Jin, L., Kaneko, O., Takeo, S., Han, E.-T., Tachibana, M., Otsuki, H., Torii, M. & Tsuboi, T., 2009. A small-scale systematic analysis of alternative splicing in Plasmodium falciparum. *Parasitology international*, 58(2), pp.196–9.

Irimia, M., Rukov, J.L., Penny, D., Garcia-Fernandez, J., Vinther, J. & Roy, S.W., 2008. Widespread evolutionary conservation of alternatively spliced exons in Caenorhabditis. *Molecular Biology and Evolution*, 25(2), pp.375–382.

Irimia, M., Rukov, J.L., Penny, D. & Roy, S.W., 2007. Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing. *BMC evolutionary biology*, 7, p.188.

Irimia, M., Rukov, J.L., Roy, S.W., Vinther, J. & Garcia-Fernandez, J., 2009. Quantitative regulation of alternative splicing in evolution and development. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 31(1), pp.40–50.

Ishii, S., Nakao, S., Minamikawa-Tachino, R., Desnick, R.J. & Fan, J.-Q., 2002. Alternative splicing in the alpha-galactosidase A gene: increased exon inclusion results in the Fabry cardiac phenotype. *American journal of human genetics*, 70(4), pp.994–1002.

Jarosch, A., Stolle, E., Crewe, R.M. & Moritz, R.F.A., 2011. Alternative splicing of a single transcription factor drives selfish reproductive behavior in honeybee workers (Apis mellifera). *Proceedings of the National Academy of Sciences of the United States of America*, 108(37), pp.15282–7.

Jerlström-Hultqvist, J., Ankarklev, J. & Svärd, S.G., Is human giardiasis caused by two different Giardia species? *Gut microbes*, 1(6), pp.379–82.

Jin, L., Kryukov, K., Clemente, J.C., Komiyama, T., Suzuki, Y., Imanishi, T., Ikeo, K. & Gojobori, T., 2008. The evolutionary relationship between gene duplication and alternative splicing. *Gene*, 427(1-2), pp.19–31.

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R. & Shoemaker, D.D., 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science (New York, N.Y.)*, 302(5653), pp.2141–4.

Johnson, N. a, 2010. Hybrid incompatibility genes: remnants of a genomic battlefield? *Trends in genetics : TIG*, 26(7), pp.317–25.

Juhász, G., Csikós, G., Sinka, R., Erdélyi, M. & Sass, M., 2003. The Drosophila homolog of Aut1 is essential for autophagy and development. *FEBS Letters*, 543(1-3), pp.154–158.

Kalsotra, A. & Cooper, T.A., 2011. Functional consequences of developmentally regulated alternative splicing. *Nature reviews. Genetics*, 12(10), pp.715–29.

Kalyna, M., Simpson, C.G., Syed, N.H., Lewandowska, D., Marquez, Y., Kusenda, B., Marshall, J., Fuller, J., Cardle, L., McNicol, J., Dinh, H.Q., Barta, A. & Brown, J.W.S., 2011. Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic acids research*, pp.1–16.

Kan, Z.Y., States, D. & Gish, W., 2002. Selecting for Functional Alternative Splices in ESTs. *Genome research*, 12(12), pp.1837–1845.

Kang, H.J., Kawasawa, Y.I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A.M.M., Pletikos, M., Meyer, K.A., Sedmak, G., Guennel, T., Shin, Y., Johnson, M.B., Krsnik, Ž., Mayer, S., Fertuzinhos, S., Umlauf, S., Lisgo, S.N., Vortmeyer, A., et al., 2011. Spatio-temporal transcriptome of the human brain. *Nature*, 478(7370), pp.483–489.

Kashima, T. & Manley, J.L., 2003. A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. *Nature genetics*, 34(4), pp.460–3.

Katsumata, T., Hosea, D., Ranuh, I.G., Uga, S., Yanagi, T. & Kohno, S., 2000. Short report: possible Cryptosporidium muris infection in humans. *The American journal of tropical medicine and hygiene*, 62(1), pp.70–2.

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M. & Stamm, S., 2013. Function of alternative splicing. *Gene*, 514(1), pp.1–30.

Kempken, F., 2013. Alternative splicing in ascomycetes. *Applied microbiology and biotechnology*, 97(10), pp.4235–41.

Keren, H., Lev-Maor, G. & Ast, G., 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nature reviews. Genetics*, 11(5), pp.345–55.

Kim, E., Goren, A. & Ast, G., 2008a. Alternative splicing: current perspectives. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 30(1), pp.38–47.

Kim, E., Goren, A. & Ast, G., 2008b. Insights into the connection between cancer and alternative splicing. *Trends in genetics : TIG*, 24(1), pp.7–10.

Kim, E., Magen, A. & Ast, G., 2007. Different levels of alternative splicing among eukaryotes. *Nucleic acids research*, 35(1), pp.125–31.

Kim, H., Klein, R., Majewski, J. & Ott, J., 2004. Estimating rates of alternative splicing in mammals and invertebrates. *Nature genetics*, 36(9), pp.915–6; author reply 916–7.

Kim, N., Alekseyenko, A. V, Roy, M. & Lee, C., 2007. The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. *Nucleic acids research*, 35(Database issue), pp.D93–8.

Kim, N. & Lee, C., 2008. Bioinformatics detection of alternative splicing. *Methods in molecular biology (Clifton, N.J.)*, 452, pp.179–97.

Kinsella, R.J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P. & Flicek, P., 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database : the journal of biological databases and curation*, 2011, p.bar030.

Kirkness, E.F., Haas, B.J., Sun, W., Braig, H.R., Perotti, M.A., Clark, J.M., Lee, S.H., Robertson, H.M., Kennedy, R.C., Elhaik, E., Gerlach, D., Kriventseva, E. V, Elsik, C.G., Graur, D., Hill, C. a, Veenstra, J. a, Walenz, B., Tubío, J.M.C., Ribeiro, J.M.C., et al., 2010. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proceedings of the National Academy of Sciences of the United States of America*, 107(27), pp.12168–73.

Kittler, R., Kayser, M. & Stoneking, M., 2003. Molecular Evolution of Pediculus humanus and the Origin of Clothing. *Current Biology*, 13(16), pp.1414–1417.

Kopelman, N.M., Lancet, D. & Yanai, I., 2005. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nature Genetics*, 37(6), pp.588–589.

Koralewski, T.E. & Krutovsky, K. V, 2011. Evolution of exon-intron structure and alternative splicing. *PloS one*, 6(3), p.e18055.

Kornblihtt, A.R., Schor, I.E., Alló, M., Dujardin, G., Petrillo, E. & Muñoz, M.J., 2013. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature reviews. Molecular cell biology*, 14(3), pp.153–65.

Koscielny, G., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Riethoven, J.J., Nardone, F., Stanley, E., Fallsehr, C., Hofmann, O., Kull, M., Harrington, E., Boue, S., Eyras, E., Plass, M., Lopez, F., Ritchie, W., Moucadel, V., Ara, T., Pospisil, H., et al., 2009. ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics*, 93(3), pp.213–220.

Krijgsheld, P., Bleichrodt, R., van Veluw, G.J., Wang, F., Müller, W.H., Dijksterhuis, J. & Wösten, H.A.B., 2013. Development in Aspergillus. *Studies in mycology*, 74(1), pp.1–29.

Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology*, 305(3), pp.567–580.

Kupfer, D.M., Drabenstot, S.D., Buchanan, K.L., Lai, H., Zhu, H., Dyer, D.W., Roe, B.A. & Murphy, J.W., 2004. Introns and splicing elements of five diverse fungi. *Eukaryotic cell*, 3(5), pp.1088–1100.

Kurtz, J. & Armitage, S. a O., 2006. Alternative adaptive immunity in invertebrates. *Trends in immunology*, 27(11), pp.493–6.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., et al., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860–921.

Lee, C.Y. & Baehrecke, E.H., 2001. Steroid regulation of autophagic programmed cell death during development. *Development (Cambridge, England)*, 128(8), pp.1443–55.

Lee, J., Lee, T., Lee, Y.-W., Yun, S.-H. & Turgeon, B.G., 2003. Shifting fungal reproductive mode by manipulation of mating type genes: obligatory heterothallism of Gibberella zeae. *Molecular Microbiology*, 50(1), pp.145–152.

Lee, Y.Y., Kim, B., Shin, Y., Nam, S., Kim, P., Kim, N., Chung, W.-H.H., Kim, J. & Lee, S., 2007. ECgene: an alternative splicing database update. *Nucleic Acids Research*, 35(Database issue), pp.D99–D103.

Lejeune, F. & Maquat, L.E., 2005. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. *Current opinion in cell biology*, 17(3), pp.309–15.

Lemaitre, B. & Hoffmann, J., 2007. The host defense of Drosophila melanogaster. *Annual review of immunology*, 25, pp.697–743.

Leo, N.P. & Barker, S.C., 2005. Unravelling the evolution of the head lice and body lice of humans. *Parasitology research*, 98(1), pp.44–7.

Leo, N.P., Campbell, N.J.H., Yang, X., Mumcuoglu, K. & Barker, S.C., 2002. Evidence from Mitochondrial DNA That Head Lice and Body Lice of Humans (Phthiraptera: Pediculidae) are Conspecific. *Journal of Medical Entomology*, 39(4), pp.662–666.

Leo, N.P., Hughes, J.M., Yang, X., Poudel, S.K.S., Brogdon, W.G. & Barker, S.C., 2005. The head and body lice of humans are genetically distinct (Insecta: Phthiraptera, Pediculidae): evidence from double infestations. *Heredity*, 95(1), pp.34–40.

Lewis, B.P., Green, R.E. & Brenner, S.E., 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 100(1), pp.189–192.

Li, C.H., Cervantes, M., Springer, D.J., Boekhout, T., Ruiz-Vazquez, R.M., Torres-Martinez, S.R., Heitman, J. & Lee, S.C., 2011. Sporangiospore size dimorphism is linked to virulence of Mucor circinelloides. *PLoS pathogens*, 7(6), p.e1002086.

Li, W., Ortiz, G., Fournier, P.-E., Gimenez, G., Reed, D.L., Pittendrigh, B. & Raoult, D., 2010. Genotyping of human lice suggests multiple emergencies of body lice from local head louse populations. *PLoS neglected tropical diseases*, 4(3), p.e641.

Light, J.E., Toups, M.A. & Reed, D.L., 2008. What's in a name: the taxonomic status of human head and body lice. *Molecular phylogenetics and evolution*, 47(3), pp.1203–16.

Lin, H., Ouyang, S., Egan, A., Nobuta, K., Haas, B.J., Zhu, W., Gu, X., Silva, J.C., Meyers, B.C. & Buell, C.R., 2008. Characterization of paralogous protein families in rice. *BMC plant biology*, 8, p.18.

Ling, K.-H., Rajandream, M.-A., Rivailler, P., Ivens, A., Yap, S.-J., Madeira, A.M.B.N., Mungall, K., Billington, K., Yee, W.-Y., Bankier, A.T., Carroll, F., Durham, A.M., Peters, N., Loo, S.-S., Isa, M.N.M., Novaes, J., Quail, M., Rosli, R., Nor Shamsudin, M., et al., 2007. Sequencing and analysis of chromosome 1 of Eimeria tenella reveals a unique segmental organization. *Genome research*, 17(3), pp.311–9.

Lister, J.A., Close, J. & Raible, D.W., 2001. Duplicate mitf genes in zebrafish: Complementary expression and conservation of melanogenic potential. *Developmental Biology*, 237(2), pp.333–344.

Liu, Y., Liu, H., Liu, S., Wang, S., Jiang, R.-J. & Li, S., 2009. Hormonal and nutritional regulation of insect fat body development and function. *Archives of insect biochemistry and physiology*, 71(1), pp.16–30.

Loftus, B.J., Fung, E., Roncaglia, P., Rowley, D., Amedeo, P., Bruno, D., Vamathevan, J., Miranda, M., Anderson, I.J., Fraser, J. a, Allen, J.E., Bosdet, I.E., Brent, M.R., Chiu, R., Doering, T.L., Donlin, M.J., D'Souza, C. a, Fox, D.S., Grinberg, V., et al., 2005. The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. *Science (New York, N.Y.)*, 307(5713), pp.1321–1324.

Long, M., Betran, E., Thornton, K. & Wang, W., 2003. The origin of new genes: Glimpses from the young and old. *Nature Reviews Genetics*, 4(11), pp.865–875.

Long, M., de Souza, S.J. & Gilbert, W., 1997. The yeast splice site revisited: new exon consensus from genomic analysis. *Cell*, 91(6), pp.739–740.

López-Bigas, N., Audit, B., Ouzounis, C., Parra, G. & Guigó, R., 2005. Are splicing mutations the most frequent cause of hereditary disease? *FEBS letters*, 579(9), pp.1900–1903.

Lowe, C.D., Mello, L. V, Samatar, N., Martin, L.E., Montagnes, D.J.S. & Watts, P.C., 2011. The transcriptome of the novel dinoflagellate Oxyrrhis marina (Alveolata: Dinophyceae): response to salinity examined by 454 sequencing. *BMC genomics*, 12(1), p.519.

Lu, F., Jiang, H., Ding, J., Mu, J., Valenzuela, J.G., Ribeiro, J.M.C. & Su, X., 2007. cDNA sequences reveal considerable gene prediction inaccuracy in the Plasmodium falciparum genome. *BMC genomics*, 8, p.255.

Luco, R.F. & Misteli, T., 2011. More than a splicing code: integrating the role of RNA, chromatin and non-coding RNA in alternative splicing regulation. *Current opinion in genetics & development*, 21(4), pp.366–72.

Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M. & Misteli, T., 2010. Regulation of alternative splicing by histone modifications. *Science*, 327(5968), pp.996–1000.

Lumsden, R.D., 1976. Ecology and Epidemiology of Pythium Species in Field Soil. *Phytopathology*, 66(10), p.1203.

Lyko, F., Foret, S., Kucharski, R., Wolf, S., Falckenhayn, C. & Maleszka, R., 2010. The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS biology*, 8(11), p.e1000506.

Lynch, M. & Conery, J.S., 2003. The origins of genome complexity. *Science*, 302(5649), pp.1401–1404.

MacLean, D.W., Meedel, T.H. & Hastings, K.E.M., 1997. Tissue-specific alternative splicing of ascidian troponin I isoforms - Redesign of a protein isoform-generating mechanism during chordate evolution. *Journal of Biological Chemistry*, 272(51), pp.32115–32120.

Malko, D.B., Makeev, V.J., Mironov, A.A. & Gelfand, M.S., 2006. Evolution of exon-intron structure and alternative splicing in fruit flies and malarial mosquito genomes. *Genome research*, 16(4), pp.505–509.

Malone, J.H. & Oliver, B., 2011. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology*, 9, p.34.

Maquat, L.E., 2004. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nature reviews. Molecular cell biology*, 5(2), pp.89–99.

Marquez, Y., Brown, J.W.S., Simpson, C., Barta, A. & Kalyna, M., 2012. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome research*, 22(6), pp.1184–95.

Marshall, A.N., Montealegre, M.C., Jiménez-López, C., Lorenz, M.C. & van Hoof, A., 2013. Alternative Splicing and Subfunctionalization Generates Functional Diversity in Fungal Proteomes J. Heitman, ed. *PLoS Genetics*, 9(3), p.e1003376.

Martin, J. a & Wang, Z., 2011. Next-generation transcriptome assembly. *Nature reviews. Genetics*, 12(10), pp.671–82.

Matlin, A.J., Clark, F. & Smith, C.W.J., 2005. Understanding alternative splicing: towards a cellular code. *Nature reviews. Molecular cell biology*, 6(5), pp.386–98.

Matthews, B.J., Kim, M.E., Flanagan, J.J., Hattori, D., Clemens, J.C., Zipursky, S.L. & Grueber, W.B., 2007. Dendrite self-avoidance is controlled by Dscam. *Cell*, 129(3), pp.593–604.

McAllister, M.M., Dubey, J.P., Lindsay, D.S., Jolley, W.R., Wills, R. a & McGuire, a M., 1998. Dogs are definitive hosts of Neospora caninum. *International journal for parasitology*, 28(9), pp.1473–8.

McGlincy, N.J. & Smith, C.W.J., 2008. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends in biochemical sciences*, 33(8), pp.385–93.

McGuire, A.M., Pearson, M.D., Neafsey, D.E. & Galagan, J.E., 2008. Cross-kingdom patterns of alternative splicing and splice recognition. *Genome biology*, 9(3), p.R50.

McQuilton, P., St Pierre, S.E. & Thurmond, J., 2012. FlyBase 101--the basics of navigating FlyBase. *Nucleic acids research*, 40(Database issue), pp.D706–14.

Mendonça, A.G., Alves, R.J. & Pereira-Leal, J.B., 2011. Loss of genetic redundancy in reductive genome evolution. *PLoS computational biology*, 7(2), p.e1001082.

Merhej, V., Royer-Carenzi, M., Pontarotti, P. & Raoult, D., 2009. Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biology direct*, 4, p.13.

Merkin, J., Russell, C., Chen, P. & Burge, C.B., 2012. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science (New York, N.Y.)*, 338(6114), pp.1593–9.

Michel, K., Budd, A., Pinto, S., Gibson, T.J. & Kafatos, F.C., 2005. Anopheles gambiae SRPN2 facilitates midgut invasion by the malaria parasite Plasmodium berghei. *EMBO reports*, 6(9), pp.891–7.

Miles, C., Elgar, G., Coles, E., Kleinjan, D.-J., van Heyningen, V. & Hastie, N., 1998. Complete sequencing of the Fugu WAGR region from WT1 to PAX6: Dramatic compaction and conservation of synteny with human chromosome 11p13. *Proceedings of the National Academy of Sciences*, 95(22), pp.13068–13072.

Miller, L.H., Baruch, D.I., Marsh, K. & Doumbo, O.K., 2002. The pathogenic basis of malaria. *Nature*, 415(6872), pp.673–9.

Mironov, A.A., Fickett, J.W. & Gelfand, M.S., 1999. Frequent alternative splicing of human genes. *Genome research*, 9(12), pp.1288–93.

Mirth, C.K. & Riddiford, L.M., 2007. Size assessment and growth control: how adult size is determined in insects. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 29(4), pp.344–55.

Modrek, B. & Lee, C.J., 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature genetics*, 34(2), pp.177–80.

Modrek, B., Resch, a, Grasso, C. & Lee, C., 2001. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic acids research*, 29(13), pp.2850–9.

Montagnes, D., Roberts, E., Lukeš, J. & Lowe, C., 2012. The rise of model protozoa. *Trends in microbiology*, 20(4), pp.184–91.

Moon, R.T., Brown, J.D. & Torres, M., 1997. WNTs modulate cell fate and behavior during vertebrate development. *Trends in genetics : TIG*, 13(4), pp.157–62.

Moore, M.J. & Silver, P.A., 2008. Global analysis of mRNA splicing. *RNA (New York, N.Y.)*, 14(2), pp.197–203.

Moore, R.C. & Purugganan, M.D., 2005. The evolutionary dynamics of plant duplicate genes. *Current opinion in plant biology*, 8(2), pp.122–8.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7), pp.621–628.

Moseley, C.T., 2002. An Exon Splice Enhancer Mutation Causes Autosomal Dominant GH Deficiency. *Journal of Clinical Endocrinology & Metabolism*, 87(2), pp.847–852.

Mudge, J.M., Frankish, A., Fernandez-Banet, J., Alioto, T., Derrien, T., Howald, C., Reymond, A., Guigo, R., Hubbard, T. & Harrow, J., 2011. The origins, evolution and functional potential of alternative splicing in vertebrates. *Molecular Biology and Evolution*, (28), pp.2949–2959.

Muhia, D.K., Swales, C.A., Eckstein-Ludwig, U., Saran, S., Polley, S.D., Kelly, J.M., Schaap, P., Krishna, S. & Baker, D.A., 2003. Multiple splice variants encode a novel adenylyl cyclase of possible plastid origin expressed in the sexual stage of the malaria parasite Plasmodium falciparum. *The Journal of biological chemistry*, 278(24), pp.22014–22.

Mullen, G.R. & Durden, L.A., 2002. *Medical and Veterinary Entomology*, Academic Press Inc.

Nagaraj, S.H., Gasser, R.B. & Ranganathan, S., 2007. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Briefings in bioinformatics*, 8(1), pp.6–21.

Nakai, K. & Horton, P., 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in Biochemical Sciences*, 24(1), pp.34–35.

Ner-Gaon, H., Halachmi, R., Savaldi-Goldstein, S., Rubin, E., Ophir, R. & Fluhr, R., 2004. Intron retention is a major phenomenon in alternative splicing in Arabidopsis. *The Plant journal : for cell and molecular biology*, 39(6), pp.877–85.

Nezis, I.P., Lamark, T., Velentzas, A.D., Rusten, T.E., Bjørkøy, G., Johansen, T., Papassideri, I.S., Stravopodis, D.J., Margaritis, L.H., Stenmark, H. & Brech, A., 2009. Cell death during Drosophila melanogaster early oogenesis is mediated through autophagy. *Autophagy*, 5(3), pp.298–302.

Ni, J.Z., Grate, L., Donohue, J.P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T. a, Blume, J.E. & Ares, M., 2007. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes & development*, 21(6), pp.708–18.

Nilsen, T.W. & Graveley, B.R., 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280), pp.457–463.

Nuttall, G.H.F., 1940. The Louse: An Account of the Lice Which Infest Man, Their Medical Importance and Control. *Journal of the American Medical Association*, 115(17), p.1479.

Ober, D., 2005. Seeing double: gene duplication and diversification in plant secondary metabolism. *Trends in plant science*, 10(9), pp.444–9.

Olds, B.P., Coates, B.S., Steele, L.D., Sun, W., Agunbiade, T. a, Yoon, K.S., Strycharz, J.P., Lee, S.H., Paige, K.N., Clark, J.M. & Pittendrigh, B.R., 2012. Comparison of the transcriptional profiles of head and body lice. *Insect molecular biology*, 21(2), pp.257–68.

Opperman, C.H., Bird, D.M., Williamson, V.M., Rokhsar, D.S., Burke, M., Cohn, J., Cromer, J., Diener, S., Gajan, J., Graham, S., Houfek, T.D., Liu, Q., Mitros, T., Schaff, J., Schaffer, R., Scholl, E., Sosinski, B.R., Thomas, V.P. & Windham, E., 2008. Sequence and genetic map of Meloidogyne hapla: A compact nematode genome for plant parasitism. *Proceedings of the National Academy of Sciences of the United States of America*, 105(39), pp.14802–7.

Orengo, J.P. & Cooper, T.A., 2007. Alternative splicing in disease. *Advances in experimental medicine and biology*, 623, pp.212–23.

Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N. & Pearse, W., 2012. caper: Comparative Analyses of Phylogenetics and Evolution in R.

Osmark, P., Hansson, O., Jonsson, A., Rönn, T., Groop, L. & Renström, E., 2009. Unique splicing pattern of the TCF7L2 gene in human pancreatic islets. *Diabetologia*, 52(5), pp.850–4.

Ozsolak, F. & Milos, P.M., 2011. RNA sequencing: advances, challenges and opportunities. *Nature reviews. Genetics*, 12(2), pp.87–98.

Pagel, M., 1999. Inferring the historical patterns of biological evolution. *Nature*, 401(6756), pp.877–84.

Pain, A., Renauld, H., Berriman, M., Murphy, L., Yeats, C.A., Weir, W., Kerhornou, A., Aslett, M., Bishop, R., Bouchier, C., Cochet, M., Coulson, R.M.R., Cronin, A., de Villiers, E.P., Fraser, A., Fosker, N., Gardner, M., Goble, A., Griffiths-Jones, S., et al., 2005. Genome of the host-cell transforming parasite Theileria annulata compared with T. parva. *Science (New York, N.Y.)*, 309(5731), pp.131–3.

Pajares, M.J., Ezponda, T., Catena, R., Calvo, A., Pio, R. & Montuenga, L.M., 2007. Alternative splicing: an emerging topic in molecular and clinical oncology. *The lancet oncology*, 8(4), pp.349–57.

Palmer, C.J., Xiao, L., Terashima, A., Guerra, H., Gotuzzo, E., Saldías, G., Bonilla, J.A., Zhou, L., Lindquist, A. & Upton, S.J., 2003. Cryptosporidium muris, a rodent pathogen, recovered from a human in Perú. *Emerging infectious diseases*, 9(9), pp.1174–6.

Pan, Q., Bakowski, M.A., Morris, Q., Zhang, W., Frey, B.J., Hughes, T.R. & Blencowe, B.J., 2005. Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends in genetics : TIG*, 21(2), pp.73–7.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J., 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12), pp.1413–1415.

Paterson, A. & Gray, R., 1997. Host-parasite co-speciation, host switching, and missing the boat. In *Host-parasite evolution: general principles and avian models.* pp. 236–250.

Peifer, M., 2000. Wnt Signaling in Oncogenesis and Embryogenesis--a Look Outside the Nucleus. *Science*, 287(5458), pp.1606–1609.

Pham Duc, P., Nguyen-Viet, H., Hattendorf, J., Zinsstag, J., Dac Cam, P. & Odermatt, P., 2011. Risk factors for Entamoeba histolytica infection in an agricultural community in Hanam province, Vietnam. *Parasites & vectors*, 4, p.102.

Pickrell, J.K., Pai, A.A., Gilad, Y. & Pritchard, J.K., 2010. Noisy splicing drives mRNA isoform diversity in human cells. *Plos Genetics*, 6(12).

Pittendrigh, B.R., Clark, J.M., Johnston, J.S., Lee, S.H., Romero-Severson, J. & Dasch, G.A., 2006. Sequencing of a new target genome: the Pediculus humanus

humanus (Phthiraptera: Pediculidae) genome project. *Journal of medical entomology*, 43(6), pp.1103–11.

Prokunina-Olsson, L., Welch, C., Hansson, O., Adhikari, N., Scott, L.J., Usher, N., Tong, M., Sprau, A., Swift, A., Bonnycastle, L.L., Erdos, M.R., He, Z., Saxena, R., Harmon, B., Kotova, O., Hoffman, E.P., Altshuler, D., Groop, L., Boehnke, M., et al., 2009. Tissue-specific alternative splicing of TCF7L2. *Human molecular genetics*, 18(20), pp.3795–804.

Ramani, A.K., Calarco, J. a, Pan, Q., Mavandadi, S., Wang, Y., Nelson, A.C., Lee, L.J., Morris, Q., Blencowe, B.J., Zhen, M. & Fraser, A.G., 2011. Genome-wide analysis of alternative splicing in Caenorhabditis elegans. *Genome research*, 21(2), pp.342–8.

Ramirez, N.E., Ward, L. a & Sreevatsan, S., 2004. A review of the biology and epidemiology of cryptosporidiosis in humans and animals. *Microbes and infection / Institut Pasteur*, 6(8), pp.773–85.

Reddy, A.S.N., 2007. Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annual review of plant biology*, 58, pp.267–94.

Reed, D.L., Light, J.E., Allen, J.M. & Kirchman, J.J., 2007. Pair of lice lost or parasites regained: the evolutionary history of anthropoid primate lice. *BMC biology*, 5, p.7.

Reed, D.L., Smith, V.S., Hammond, S.L., Rogers, A.R. & Clayton, D.H., 2004. Genetic analysis of lice supports direct contact between modern and archaic humans. *PLoS biology*, 2(11), p.e340.

Dos Reis, M., Inoue, J., Hasegawa, M., Asher, R.J., Donoghue, P.C.J. & Yang, Z., 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proceedings. Biological sciences / The Royal Society*, 279(1742), pp.3491–500.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y.S., Newsome, R., Chan, S.K., She, R., Varhol, R., et al., 2010. De novo assembly and analysis of RNA-seq data. *Nature methods*, 7(11), pp.909–912.

Roux, J. & Robinson-Rechavi, M., 2011. Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. *Genome research*, 21(3), pp.357–63.

Rusten, T.E., Lindmo, K., Juhász, G., Sass, M., Seglen, P.O., Brech, A. & Stenmark, H., 2004. Programmed autophagy in the Drosophila fat body is induced by

ecdysone through regulation of the PI3K pathway. *Developmental cell*, 7(2), pp.179–92.

Rydkina, E.B., Roux, V., Gagua, E.M., Predtechenski, a B., Tarasevich, I. V & Raoult, D., 1999. Bartonella quintana in body lice collected from homeless persons in Russia. *Emerging infectious diseases*, 5(1), pp.176–8.

Salz, H.K., 2011. Sex determination in insects: a binary decision based on alternative splicing. *Current opinion in genetics & development*, 21(4), pp.395–400.

Sasaki-Fukatsu, K., Koga, R., Nikoh, N., Yoshizawa, K., Kasai, S., Mihara, M., Kobayashi, M., Tomita, T. & Fukatsu, T., 2006. Symbiotic bacteria associated with stomach discs of human lice. *Applied and environmental microbiology*, 72(11), pp.7349–52.

Saxena, R., Gianniny, L., Burtt, N.P., Lyssenko, V., Giuducci, C., Sjögren, M., Florez, J.C., Almgren, P., Isomaa, B., Orho-Melander, M., Lindblad, U., Daly, M.J., Tuomi, T., Hirschhorn, J.N., Ardlie, K.G., Groop, L.C. & Altshuler, D., 2006. Common single nucleotide polymorphisms in TCF7L2 are reproducibly associated with type 2 diabetes and reduce the insulin response to glucose in nondiabetic individuals. *Diabetes*, 55(10), pp.2890–5.

Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Miller, V., et al., 2009. Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 37(Database issue), pp.D5–15.

Scala, S., Carels, N., Falciatore, A., Chiusano, M.L. & Bowler, C., 2002. Genome properties of the diatom Phaeodactylum tricornutum. *Plant physiology*, 129(3), pp.993–1002.

Schad, E., Tompa, P. & Hegyi, H., 2011. The relationship between proteome size, structural disorder and organism complexity. *Genome biology*, 12(12), p.R120.

Schaefer, C.W., 1978. Ecological separation of the human head lice and body lice (Anoplura: Pediculidae). *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 72(6), pp.669–70.

Schindler, S., Szafranski, K., Hiller, M., Ali, G.S., Palusa, S.G., Backofen, R., Platzer, M. & Reddy, A.S.N., 2008. Alternative splicing at NAGNAG acceptors in Arabidopsis thaliana SR and SR-related protein-coding genes. *BMC genomics*, 9, p.159.

Schmucker, D., Clemens, J.C., Shu, H., Worby, C.A., Xiao, J., Muda, M., Dixon, J.E. & Zipursky, S.L., 2000. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, 101(6), pp.671–84.

Schwartz, S.H., Silva, J., Burstein, D., Pupko, T., Eyras, E. & Ast, G., 2008. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome research*, 18(1), pp.88–103.

Shattock, R.C., 2002. Phytophthora infestans: populations, pathogenicity and phenylamides. *Pest management science*, 58(9), pp.944–50.

Short, S. & Holland, L.Z., 2008. The evolution of alternative splicing in the Pax family: The view from the basal chordate amphioxus. *Journal of Molecular Evolution*, 66(6), pp.605–620.

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R. & Oberdoerffer, S., 2011. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, 479(7371), pp.74–79.

Singh, N., Preiser, P., Rénia, L., Balu, B., Barnwell, J., Blair, P., Jarra, W., Voza, T., Landau, I., Adams, J.H., Iriko, H., Jin, L., Kaneko, O., Takeo, S., Han, E.-T., Tachibana, M., Otsuki, H., Torii, M. & Tsuboi, T., 2004. Conservation and developmental control of alternative splicing in maebl among malaria parasites. *Parasitology international*, 58(3), pp.196–9.

Singh, S., Mishra, R., Arango, N.A., Deng, J.M., Behringer, R.R. & Saunders, G.F., 2002. Iris hypoplasia in mice that lack the alternatively spliced Pax6(5a) isoform. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10), pp.6812–6815.

Skotheim, R.I. & Nees, M., 2007. Alternative splicing in cancer: noise, functional, or systematic? *The international journal of biochemistry & cell biology*, 39(7-8), pp.1432–1449.

Smith, P.H., Mwangi, J.M., Afrane, Y. a, Yan, G., Obbard, D.J., Ranford-Cartwright, L.C. & Little, T.J., 2011. Alternative splicing of the Anopheles gambiae Dscam gene in diverse Plasmodium falciparum infections. *Malaria journal*, 10(1), p.156.

Sorek, R., Shamir, R. & Ast, G., 2004. How prevalent is functional alternative splicing in the human genome? *Trends in genetics : TIG*, 20(2), pp.68–71.

Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T. a & Soreq, H., 2005. Function of alternative splicing. *Gene*, 344, pp.1–20.

Stuart, L.M., Boulais, J., Charriere, G.M., Hennessy, E.J., Brunet, S., Jutras, I., Goyette, G., Rondeau, C., Letarte, S., Huang, H., Ye, P., Morales, F., Kocks, C., Bader, J.S., Desjardins, M. & Ezekowitz, R. a B., 2007. A systems biology analysis of the Drosophila phagosome. *Nature*, 445(7123), pp.95–101.

Sturm, A., Amino, R., van de Sand, C., Regen, T., Retzlaff, S., Rennenberg, A., Krueger, A., Pollok, J.-M., Menard, R. & Heussler, V.T., 2006. Manipulation of host hepatocytes by the malaria parasite for delivery into liver sinusoids. *Science (New York, N.Y.)*, 313(5791), pp.1287–90.

Su, Z. & Gu, X., 2012. Revisit on the evolutionary relationship between alternative splicing and gene duplication. *Gene*, 504(1), pp.102–6.

Su, Z., Wang, J., Yu, J., Huang, X. & Gu, X., 2006. Evolution of alternative splicing after gene duplication. *Genome research*, 16(2), pp.182–9.

Sugnet, C.W., Kent, W.J., Ares, M. & Haussler, D., 2004. Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 77, pp.66–77.

Sun, S.H. & Huppert, M., 1976. A cytological study of morphogenesis in Coccidioides immitis. *Medical Mycology*, 14(2), pp.185–198.

Swank, D.M., Knowles, A.F., Suggs, J.A., Sarsoza, F., Lee, A., Maughan, D.W. & Bernstein, S.I., 2002. The myosin converter domain modulates muscle performance. *Nature cell biology*, 4(4), pp.312–6.

Talavera, D., Vogel, C., Orozco, M., Teichmann, S.A. & de la Cruz, X., 2007. The (In) dependence of alternative splicing and gene duplication. *Plos Computational Biology*, 3(3), pp.375–388.

Taliaferro, J.M., Alvarez, N., Green, R.E., Blanchette, M. & Rio, D.C., 2011. Evolution of a tissue-specific splicing network. *Genes & development*, 25(6), pp.608–20.

Tang, J.-Y., Lee, J.-C., Hou, M.-F., Wang, C.-L., Chen, C.-C., Huang, H.-W. & Chang, H.-W., 2013. Alternative splicing for diseases, cancers, drugs, and databases. *TheScientificWorldJournal*, 2013, p.703568.

Taylor, J.W. & Berbee, M.L., 2006. Dating divergences in the Fungal Tree of Life: review and new analyses. *Mycologia*, 98(6), pp.838–49.

Taylor, S., Barragan, A., Su, C., Fux, B., Fentress, S.J., Tang, K., Beatty, W.L., Hajj, H. El, Jerome, M., Behnke, M.S., White, M., Wootton, J.C. & Sibley, L.D., 2006. A secreted serine-threonine kinase determines virulence in the eukaryotic

pathogen Toxoplasma gondii. *Science (New York, N.Y.)*, 314(5806), pp.1776–80.

Tazi, J., Bakkour, N. & Stamm, S., 2009. Alternative splicing and disease. *Biochimica et biophysica acta*, 1792(1), pp.14–26.

Thanaraj, T.A., Clark, F. & Mulilu, J., 2003. Conservation of human alternative splice events in mouse. *Nucleic Acids Research*, 31, pp.2544–2552.

Thomas, C.A., 1971. The genetic organization of chromosomes. *Annual review of genetics*, 5, pp.237–56.

Toups, M.A., Kitchen, A., Light, J.E. & Reed, D.L., 2011. Origin of clothing lice indicates early clothing use by anatomically modern humans in Africa. *Molecular biology and evolution*, 28(1), pp.29–32.

Truman, J.W., Talbot, W.S., Fahrbach, S.E. & Hogness, D.S., 1994. Ecdysone receptor expression in the CNS correlates with stage-specific responses to ecdysteroids during Drosophila and Manduca development. *Development (Cambridge, England)*, 120(1), pp.219–34.

Tyler, B.M., 2007. Phytophthora sojae: root rot pathogen of soybean and model oomycete. *Molecular plant pathology*, 8(1), pp.1–8.

Uv, A.E., Harrison, E.J. & Bray, S.J., 1997. Tissue-specific splicing and functions of the Drosophila transcription factor Grainyhead. *Molecular and cellular biology*, 17(11), pp.6727–35.

Venables, J.P., 2006. Unbalanced alternative splicing and its significance in cancer. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 28(4), pp.378–386.

Venables, J.P., Tazi, J. & Juge, F., 2012. Regulated functional alternative splicing in Drosophila. *Nucleic acids research*, 40(1), pp.1–10.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., et al., 2001. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507), pp.1304–51.

Veracx, A., Boutellis, A., Merhej, V., Diatta, G. & Raoult, D., 2012. Evidence for an African cluster of human head and body lice with variable colors and interbreeding of lice between continents. *PloS one*, 7(5), p.e37804.

Wang, B.-B. & Brendel, V., 2006. Genomewide comparative analysis of alternative splicing in plants. *Proceedings of the National Academy of Sciences of the United States of America*, 103(18), pp.7175–80.

Wang, E.T., Sandberg, R., Luo, S.J., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P. & Burge, C.B., 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221), pp.470–6.

Wang, G.-S.S. & Cooper, T. a, 2007. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature reviews. Genetics*, 8(10), pp.749–61.

Wang, Z. & Burge, C.B., 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA (New York, N.Y.)*, 14(5), pp.802–13.

Wang, Z., Gerstein, M. & Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1), pp.57–63.

Wasbrough, E.R., Dorus, S., Hester, S., Howard-Murkin, J., Lilley, K., Wilkin, E., Polpitiya, A., Petritis, K. & Karr, T.L., 2010. The Drosophila melanogaster sperm proteome-II (DmSP-II). *Journal of proteomics*, 73(11), pp.2171–85.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., et al., 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915), pp.520–62.

Watson, F.L., Püttmann-Holgado, R., Thomas, F., Lamar, D.L., Hughes, M., Kondo, M., Rebel, V.I. & Schmucker, D., 2005. Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science*, 309(5742), pp.1874–8.

Waugh, M., 2000. The Phytophthora Genome Initiative Database: informatics and analysis for distributed pathogenomic research. *Nucleic Acids Research*, 28(1), pp.87–90.

Weiss, R.A., 2009. Apes, lice and prehistory. *Journal of biology*, 8(2), p.20.

Whitney, K.D. & Garland, T., 2010. Did genetic drift drive increases in genome complexity? *PLoS genetics*, 6(8).

Wu, T.D. & Watanabe, C.K., 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics (Oxford, England)*, 21(9), pp.1859–75.

Xiong, J., Lu, X., Zhou, Zhemin, Chang, Y., Yuan, D., Tian, M., Zhou, Zhigang, Wang, L., Fu, C., Orias, E. & Miao, W., 2012. Transcriptome analysis of the

model protozoan, Tetrahymena thermophila, using Deep RNA sequencing. *PloS one*, 7(2), p.e30630.

Xu, Q., 2003. Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Research*, 31(19), pp.5635–5643.

Yanagida, M., 2002. The model unicellular eukaryote, Schizosaccharomyces pombe. *Genome biology*, 3(3), p.COMMENT2003.

Yeo, G., Holste, D., Kreiman, G. & Burge, C.B., 2004. Variation in alternative splicing across human tissues. *Genome biology*, 5(10), p.R74.

Yeo, G.W., Van Nostrand, E., Holste, D., Poggio, T. & Burge, C.B., 2005. Identification and analysis of alternative splicing events conserved in human and mouse. *Proceedings of the National Academy of Sciences of the United States of America*, 102(8), pp.2850–5.

Yi, F., Brubaker, P.L. & Jin, T., 2005. TCF-4 mediates cell type-specific regulation of proglucagon gene expression by beta-catenin and glycogen synthase kinase-3beta. *The Journal of biological chemistry*, 280(2), pp.1457–64.

Yu, W.P., Brenner, S. & Venkatesh, B., 2003. Duplication, degeneration and subfunctionalization of the nested synapsin-Timp genes in Fugu. *Trends in Genetics*, 19(4), pp.180–183.

Yu, Y., Maroney, P.A., Denker, J.A., Zhang, X.H., Dybkov, O., Luhrmann, R., Jankowsky, E., Chasin, L.A. & Nilsen, T.W., 2008. Dynamic regulation of alternative splicing by silencers that modulate 5' splice site competition. *Cell*, 135(7), pp.1224–1236.

Zdobnov, E.M. & Apweiler, R., 2001. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17(9), pp.847–848.

Zhang, Z., Xin, D., Wang, P., Zhou, L., Hu, L., Kong, X. & Hurst, L.D., 2009. Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay. *Bmc Biology*, 7, p.23.

Zhao, C., Waalwijk, C., de Wit, P.J.G.M., Tang, D. & van der Lee, T., 2013. RNA-Seq analysis reveals new gene models and alternative splicing in the fungal pathogen Fusarium graminearum. *BMC genomics*, 14(1), p.21.

Zhu, W., 2003. Identification, characterization and molecular phylogeny of U12-dependent introns in the Arabidopsis thaliana genome. *Nucleic Acids Research*, 31(15), pp.4561–4572.

Zhu, W., Schlueter, S.D. & Brendel, V., 2003. Refined annotation of the Arabidopsis genome by complete expressed sequence tag mapping. *Plant physiology*, 132(2), pp.469–84.

8 Appendix

8.1 Section 1.

Chen, L., Tovar-Corona, J. M. & Urrutia, A.O., 2011. Increased levels of noisy splicing in cancers, but not for oncogene-derived transcripts. *Human molecular genetics*, 20(22), pp.4422–4429.

8.2 Section 1.

Wu, X., Tronholm, A., Fernández-Cáceres, E., Tovar-Corona, J. M., Chen, L., Urrutia, A. O., Hurst, L. D., 2013. Evidence for deep phylogenetic conservation of exonic splice-related constraints: splice-related skews at exonic ends in the brown alga Ectocarpus are common and resemble those seen in humans. *Genome biology and evolution*, pp.1–38.

8.3 Section 1.

Bush, S. J., Castillo-Morales, A., Tovar-Corona, J. M., Chen, L., Kover, P. X, Urrutia, A. O., 2013. Presence/absence variation in A. thaliana is primarily associated with genomic signatures consistent with relaxed selective constraints. *Molecular biology and evolution*, pp.1–33.

HMG Advance Access published August 23, 2011

Increased levels of noisy splicing in cancers, but not for oncogene-derived transcripts

Lu Chen[1], Jaime M. Tovar-Corona[1] and Araxi O. Urrutia[1,]*

[1]Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK

* To whom correspondence should be addressed. Tel: +44 1225386318; Fax: +44 1225386779; Email: A.Urrutia@bath.ac.uk

ABSTRACT

Recent genome-wide analyses have detected numerous cancer-specific alternative splicing (AS) events. Whether transcripts containing cancer-specific AS events are likely to be translated into functional proteins or simply reflect noisy splicing, thereby determining their clinical relevance, is not known. Here we show that consistent with a noisy-splicing model, cancer-specific AS events generally tend to be rare, containing more premature stop codons and have less identifiable functional domains in both human and mouse. Interestingly, common cancer-derived AS transcripts from tumour suppressor and oncogenes show marked changes in premature stop codon frequency; with tumour suppressor genes exhibiting increased levels of premature stop codons whereas oncogenes have the opposite pattern. We conclude that tumours tend to have faithful oncogene splicing and a higher incidence of premature stop codons among tumour suppressor and cancer-specific splice variants showing the importance of considering splicing noise when analysing cancer-specific splicing changes.

INTRODUCTION

Cancer cells are associated with profound changes at the transcriptome level with hundreds of genes being up or down regulated when compared to normal tissues (1). Transcription profiling of cancer samples has led to an increased understanding of cancer physiology and the identification of a number of transcriptional cancer markers. Alternative splicing (AS) is a post-transcriptional process in eukaryotic organisms by which multiple distinct functional transcripts are produced from a single gene. It is now known that most human genes undergo alternative splicing (2, 3). Several studies have explored cancer related changes in alternative splicing patterns (reviewed in (4-7)) resulting in the identification of an increasing number of cancer-specific AS events in a variety of cancer tissues (8-12). Given the high number of AS events unique to cancer transcriptomes, cancer-specific transcripts have been proposed to play a key role in cancer physiology (6, 12). Nevertheless, only a handful of cancer-specific alternative splicing events have been experimentally validated (8, 10). Given that a significant proportion of alternatively spliced transcripts result from noisy splicing in normal human tissues (13-16), it is possible that most cancer-specific AS result from aberrant splicing in these abnormal cells and not play any significant role in cancer onset or progression (6, 9, 11). Here, by examining human and mouse EST libraries we ask whether cancer transcriptomes show any differences in transcript quality compared to normal tissues.

RESULTS

Identification of cancer-specific alternative splicing events in human and mouse

A total of 10,896,836 ESTs for human and mouse were downloaded from UniGene (17). Of these 3,384,826 ESTs had a clear disease state annotation and were split into 313 libraries representing normal 41 tissues and 368 cancer libraries for 35 tissues for human and 192 normal libraries corresponding to 32 normal tissues and 52 cancer libraries from 16 tissues for mouse (see Table 1). To identify alternative splicing events, a complete exon template was constructed for each gene by mapping all partial and full transcripts available (using Gmap software (18)). Known nested genes as well as orphan exons, not present in any transcript extending beyond them, were removed from further analysis. Individual ESTs were then aligned to the resulting gene template to identify AS events. We identified a total of 1,349,341 and 271,491 AS transcripts containing AS events for human and mouse respectively. Of these, a total of 1,259,641 (93.3%) and 199,943 (73.6%) for human and mouse respectively were found in both normal and cancer libraries while 23,876 (1.8%) and 1,956 (0.7%) were found only in cancer libraries. The remainder 65,824 (4.9%) and 69,592 (25.6%) transcripts were found to contain AS events exclusive to normal tissue derived libraries (Figure 1). The higher percentage of normal-specific AS events in mouse is explained by the limited cancer transcripts available for this species (Table 1).

Cancer transcripts show signatures consistent with splicing noise

We then assessed whether cancer libraries and in particular cancer-specific transcripts show signatures consistent with increased rates of splicing noise. If so, we expect cancer transcripts to: A) have a higher incidence of nonsense or frameshift mutations

which introduce a premature translation termination codons to mRNAs resulting in truncated proteins or more often rendering them vulnerable to nonsense mediated decay (13, 14). In the case of cancer-specific transcripts we can expect them to: B) have reduced identifiable functional components consistent with higher rates of aberrant incorporation of non coding regions into the transcript (see methods); C) be found mostly as single copy and D) be present in only one library thus not being part of the core cancer transcription profile as these are more likely to result from splicing errors (15).

Transcripts were classified according to whether they contained AS events found in both normal and cancer tissues or unique to either resulting in four groups: 1) *Normal common*, with transcripts containing AS events also found in at least one cancer library, 2) *Normal-specific*, whose AS events are only found in normal tissue samples, 3) *Cancer common*, containing transcripts from cancer libraries with AS events also found in at least one normal tissue library and 4) Cancer-specific with transcripts with AS events unique to cancer libraries. Our results show, compared to normal tissue derived transcripts, an increased incidence of premature stop codons among cancer-derived transcripts which is higher for cancer-specific transcripts (Figure 2, $P < 0.0001$) in both human and mouse. In both species, cancer-specific events were also found to have a significantly lower number of identifiable functional components ($P < 0.0001$; Figure 3). In addition, we found that the vast majority (79.0%) have been sequenced only once with 90.5% identified in a single EST library in human (Figure 4). In contrast, normal-specific transcripts show less pronounced differences in premature stop codons and functional components compared to transcripts with normal-common AS events (Figure 2 and Figure 3). We also found that transcripts

containing AS events particular to normal tissues are significantly less likely to be found as a single copy or confined to a single library (P <= 0.0001; Figure 4).

Tumour suppressor and oncogenes reveal contrasting transcript quality reductions in cancer libraries

Because tumour suppressor and oncogenes play a key role in tumour progression, we tested whether these gene categories presented any differences in the frequencies of disabled transcripts. Inactivation of tumour suppressor genes *NFI*, *FHIT* and *TSG101* and strengthening oncogenes *CD44* and *RON* by AS have been reported (reviewed in (4, 6). To test whether splicing noise signatures affect tumour suppressor and oncogenes differently, we divided all genes into oncogenes (648), tumour suppressor (850) and *other* genes according to the CancerGenes database (19). We found that even if as a whole cancer-derived transcripts are more likely to contain premature stop codons consistent with misplicing (Figure 2), this increase is not equally distributed between gene categories (Figure 5). Common cancer-derived oncogene transcripts show only marginal changes in the rate of premature stop codons compared with transcripts derived from normal tissues (Figure 5). In contrast, tumour suppressor genes show a marked increase in the incidence of premature stop codons in cancer libraries (Figure 5, P < 0.001). These differences in transcript quality among gene categories are not observed in normal libraries.

Analyses of transcripts specific to cancer or normal tissues showed that cancer-specific AS events have an elevated rate of premature stop codons in all three categories, further suggesting that a significant proportion of cancer-specific AS events containing transcripts are likely to result from splicing errors. We also found an elevated frequency in premature stop codons among tumour suppressor derived

normal-specific AS transcripts (Figure 5; P = 0.016 and P = 0.014) which is not explained by the fact that these genes have a slightly longer average coding region (supplementary Figures 1 and 2). When comparing transcript abundance in cancer-specific AS events (Figure 6), we found that oncogenes are more likely to produce cancer-specific AS events with more than one copy and to be found in more than one library than *other* genes (P = 0.037; Figure 6). This pattern is not found for normal-specific AS transcripts were the group of *other* genes were far more likely to be present in multiple copies and multiple libraries than both tumour suppressor and oncogenes (P < 0.0001; Figure 6).

In order to assess functional content, we examined the distribution of functional components for oncogenes, tumour suppressor and other genes in both cancer and normal AS transcripts. For alternative splicing events found in both cancer and normal libraries, oncogenes and tumour suppressor derived transcripts had higher frequencies of functional components compared to *other* genes (Figure 7, P = 0.008 and P = 0.04), suggesting that alternative splicing areas contribute significantly to the functional properties of these genes protein products. While among normal-specific AS areas there is a reduction in the functional content from oncogenes; in cancer-specific AS areas, it is tumour suppressor genes which show a marked reduction in functional content. No such reduction is observed among AS areas of oncogenes (Figure 7, P = 0.019).

DISCUSSION

We have shown that transcripts derived from cancer libraries have an elevated rate of stop codons consistent with increased rates of missplicing in cancer transcriptomes. Transcripts with alternatively splicing events unique to cancer libraries showed an even greater enrichment in premature stop codons (Figure 2) as well as containing fewer identifiable functional domains (Figure 3). Importantly, all cancer-specific transcripts were found in fewer than ten cancer libraries (out of a total of 367) with almost 80% of them found as a single copy (Figure 4). These features suggest that a significant proportion of these transcripts are unlikely to produce a functional protein product and given that no cancer specific transcripts was found to be ubiquitous to all cancer libraries or even a cancer type, we believe that the majority of cancer-specific transcripts, although probably functional, are unlikely to form part of a core cancer-transcriptome. Thus we estimate that the clinical and diagnostic relevance of particular cancer-specific transcripts may prove rather limited.

In contrast, analyses of transcripts only found in normal tissue samples did not reveal a similar increase in noise signatures (Figure 2 and 3) and a significantly greater proportion were found in multiple libraries (Figure 4). Mutations leading to the absence of these transcripts in cancer libraries may have a role in cancer establishment and its progression and may therefore warrant further studies examining their clinical potential.

Interestingly, when dividing genes into oncogenes, tumour suppressors and *other* genes, we found marginal increases in stop codons in oncogene derived transcripts in cancer libraries while tumour suppressor genes showed a strong increase in premature stop codons. We found a higher incidence of premature stop codons among of tumour

suppressor genes in both normal-specific and cancer-common AS (Figure 5). This is not explained by differences in coding region length (supplementary Figures 1 and 2). The fact that cancer-specific oncogene transcripts have a higher functional content compared to those normal specific, suggests that, in some instances, oncogene-derived cancer-specific transcripts may confer novel functional properties to protein products potentially having a role in cancer cells. Given that this set of transcripts are mostly found in single libraries it is likely that their functional contribution is likely to be specific to cancers of individual patients.

We conclude that cancer states are associated with an elevated rate of aberrant transcripts particularly pronounced in tumour suppressor genes but from which oncogenes are spared. We therefore suggest that splicing noise should be considered when evaluating cancer-specific splicing events as they have a significant higher incidence of premature stop codons. Given that nonsense mutations affect only a minority of transcripts, it is feasible to assume that most cancer and normal specific transcripts may be transcribed into functional proteins and may contribute significantly to the cancerous phenotype. Nevertheless, the fact that most cancer-specific splice variants we identified are found as single copies in one EST library may somewhat limit their value as wide spectrum diagnostic probes and/or treatment targets. Assessment of global AS signatures by gene category may be more promising. Finally we propose that the roles of normal-specific and mutation in common alternative splicing variants should be examined in addition to cancer-specific transcripts; analyses of these absent AS transcripts may further aid in the understanding of the cancer physiology.

MATERIALS AND METHODS

Data sources

Sequence and genome annotations were obtained from Ensembl. EST sequences and library information were downloaded from UniGene (17).

Identification of alternative splicing events

To estimate AS events in different organisms, a novel procedure was applied as follows: (i) *Mapping predicted genes and ESTs to Genome and grouping ESTs for each gene.* Overlapping and nested genes were identified and removed from further analyses. GMAP (18) was used to align full transcripts and high quality ESTs to their corresponding predicted genes. Genes with no matching transcripts were removed from further analyses. (ii) *Template building*. To obtain a gene template as complete as possible, full transcripts and ESTs were overlaid onto the genomic sequence. This was done as follows: First the longest partial or full transcript available forms the base of the template. All other mRNAs and ESTs are then aligned to the genomic sequence and boundaries with the previously included transcripts are revised to extend exons or include new ones. If a transcript only encompasses a single exon then it will be discarded. This allows identifying any single exon which has not been previously annotated and discarding any non-supported exons annotated in “predicted gene”. (iii) *Detecting AS events.* We developed an algorithm for AS event detection to compare the exon boundaries of any transcript to its corresponding template. Discrepancies of less than 15 bp in length were discarded. To identify AS isoforms, transcripts were first sorted according to the number of AS events they contain. Then transcripts containing identical or similar AS events were classed as redundant. Each AS event was classified depending on whether it derives from cancer or normal libraries. Those

AS events not found in either normal or cancer libraries were deemed cancer or normal specific respectively, while AS events shared in both normal and cancer libraries were defined as normal common and cancer common respectively.

Identification of premature stop codons, functional and structural protein components per AS event

As transcripts supporting the same AS event may contain premature stop-codon causing mutations, stop codon presence was characterised and counted on a per transcript basis. Other features such as functional components were jointly analysed for each splicing event. To calculate the proportion of AS transcripts with stop codons, BLASTX (20) was run to search for ORF according to protein sequences. From the BLASTX alignment files, amino acid sequences of AS area were extracted and stop codons were identified and counted. To functionally characterize AS events, we used InterProScan which contains 14 applications for the prediction of protein domains (21), including Pfam for the prediction of protein domains (22), SignalP 3.0 for signal peptide predictions (23) and TMHMM (24) for the predictions of transmembrane domains. PSORT II (25) was used to identify the likely sub-cellular localization of protein products. Secondary protein structures were predicted by CLC Main Workbench 5.7, which is based on extracted protein sequences from the protein databank (http://www.rcsb.org/pdb/).

ACKNOWLEDGEMENTS

Authors wish to thank Laurence Hurst for comments on an earlier version of this manuscript. This work was funded by UK-China scholarship for excellence and University of Bath research studentship to LC, a CONACyT scholarship to JMTC and a Royal Society Dorothy Hodgkin Research Fellowship, Royal Society research grant and a Royal Society research grant for fellows to AUO.

Conflict of Interest statement. None declared.

REFERENCES

1 Martinez, O., Reyes-Valdes, M.H. and Herrera-Estrella, L. (2010) Cancer reduces transcriptome specialization. *PLoS One*, **5**, e10398.

2 Pan, Q., Shai, O., Lee, L.J., Frey, J. and Blencowe, B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413-1415.

3 Wang, E.T., Sandberg, R., Luo, S.J., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P. and Burge, C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470-476.

4 Kalnina, Z., Zayakin, P., Silina, K. and Line, A. (2005) Alterations of pre-mRNA splicing in cancer. *Genes Chromosomes Canc.*, **42**, 342-357.

5 Venables, J.P. (2006) Unbalanced alternative splicing and its significance in cancer. *BioEssays*, **28**, 378-386.

6 Skotheim, R.I. and Nees, M. (2007) Alternative splicing in cancer: noise, functional, or systematic? *Int. J. Biochem. Cell Biol.* **39**, 1432-1449.

7 Wang, G.S. and Cooper, T.A. (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.*, **8**, 749-761.

8 Wang, Z., Lo, H.S., Yang, H., Gere, S., Hu, Y., Buetow, K.H. and Lee, M.P. (2003) Computational Analysis and Experimental Validation of Tumor-associated Alternative RNA Splicing in Human Cancer Computational Analysis and Experimental Validation of Tumor-associated Alternative RNA Splicing in Human Cancer. *Cancer Res.*, 655-657.

9 Xu, Q. (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res.*, **31**, 5635-5643.

10 Hui, L., Zhang, X., Wu, X., Lin, Z., Wang, Q., Li, Y. and Hu, G. (2004) Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. *Oncogene*, **23**, 3013-3023.

11 Kim, E., Goren, A. and Ast, G. (2008) Insights into the connection between cancer and alternative splicing. *Trends genet.*, **24**, 7-10.

12 He, C., Zhou, F., Zuo, Z., Cheng, H. and Zhou, R. (2009) A global view of cancer-specific transcript variants by subtractive transcriptome-wide analysis. *PloS one*, **4**, e4732-e4732.

13 Green, R.E., Lewis, B.P., Hillman, R.T., Blanchette, M., Lareau, L.F., Garnett, A.T., Rio, D.C. and Brenner, S.E. (2003) Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. *Bioinformatics*, **19**, i118-i121.

14 Lewis, B.P., Green, R.E. and Brenner, S.E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl Acad. Sci. USA*, **100**, 189-192.

15 Zhang, Z.G., Xin, D.D., Wang, P., Zhou, L., Hu, L.D., Kong, X.Y. and Hurst, L.D. (2009) Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay. *BMC Biol.*, **7**, 13.

16 Pickrell, J.K., Pai, A.A., Gilad, Y. and Pritchard, J.K. (2010) Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.*, **6**.

17 Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Federhen, S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **38**, D5-D16.

18 Wu, T.D. and Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859-1875.

19 Higgins, M.E., Claremont, M., Major, J.E., Sander, C. and Lash, A.E. (2007) CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Res.*, **35**, D721-726.

20 Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389-3402.

21 Zdobnov, E.M. and Apweiler, R. (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847-848.

22 Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138-D141.

23 Bendtsen, J.D., Nielsen, H., von Heijne, G. and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783-795.

24 Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.*, **305**, 567-580.

25 Nakai, K. and Horton, P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, **24**, 34-35.

FIGURE LEGENDS

Figure 1. Schematic representation of the proportion of transcripts containing alternative splicing events common in both normal and cancer libraries, or cancer/normal specific. First number in each label represents the total number of distinct AS events detected and the second the number of genes represented for human (Hs) and mouse (Mm).

Figure 2. Premature stop codons in normal and cancer AS events. Top panel shows the percentage of premature stop codon containing AS events for normal and cancer tissues subdivided into those containing AS events unique to normal / cancer libraries or found in both. Bottom panel shows average number of premature stop codons with events divided in the same way as top panel. Stars represent significant differences among groups from top panels (Chi-square test) and bottom panels (Wilcoxon test) with $0.01 < P < 0.05$ (*), $0.001 < P < 0.01$ (**) and $P <= 0.001$ (***).

Figure 3. Identifiable functional components in AS events in cancer and normal tissues. Top panel shows the percentage of AS events with at least one identifiable functional component (see methods). Bottom panel shows average number of identifiable functional components per AS area. In both panels transcripts were divided as in Figure 2. Stars represent significant differences among groups from top panels (Chi-square test) and bottom panels (Wilcoxon test) with $0.01 < P <= 0.05$ (*), $0.001 < P <= 0.01$ (**) and $P <= 0.001$ (***).

Figure 4. Normal and cancer specific AS events frequency distributions. Top panel shows the number of times each AS event is found and bottom panel shows the number of libraries where an AS event is found. Error bars in distributions from normal specific transcripts represent one hundred randomly selected samples from

normal-specific transcripts of equal transcript and library number to the number of cancer-specific transcripts and libraries available.

Figure 5. Premature stop codon frequency in oncogenes, tumour suppressor and *other* genes. Top panel shows the percentage of premature stop codon containing AS events. Bottom panel shows the average number of stop codons per AS events. AS events were classified depending on whether they were derived from oncogenes tumour suppressor and *other* genes. Broader groupings from Figure 2 and Figure 3 are also labelled. Stars represent significant differences among groups from top panels (Chi-square test) and bottom panels (Wilcoxon test) with $0.01 < P < 0.05$ (*), $0.001 < P < 0.01$ (**) and $P <= 0.001$ (***).

Figure 6. AS event frequency for normal and cancer transcripts divided into oncogene tumour suppressor and *other* genes. Left and right panels represent cancer-specific and normal-specific AS events, respectively. Distributions for normal-specific AS events are the average results from 100 randomly selected samples of equal size to the number of cancer-specific AS events. Top panels present the percentage of AS events which are present in more than one copy and or more than one library. Bottom panels are a box plot of the average number of copies per AS event or the number of libraries where each AS event is present. Stars represent significant differences among groups from top panels (Chi-square test) and bottom panels (Wilcoxon test) with $0.01 < P < 0.05$ (*), $0.001 < P < 0.01$ (**) and $P <= 0.001$ (***).

Figure 7. Identifiable functional components in AS events in cancer and normal transcripts divided into oncogene, tumour suppressor and *other* gene-derived. Top panel shows the percentage of AS events with at least one identifiable functional component (see methods). Bottom panel shows average number of identifiable

functional components per AS area. In both panels, AS events were divided into groups as in Figure 3 and further subdivided into oncogene, tumour suppressor and *other* genes. Stars represent significant differences among groups from top panels (Chi-square test) and bottom panels (Wilcoxon test) with $0.01 < P < 0.05$ (*), $0.001 < P < 0.01$ (**) and $P <= 0.001$ (***).

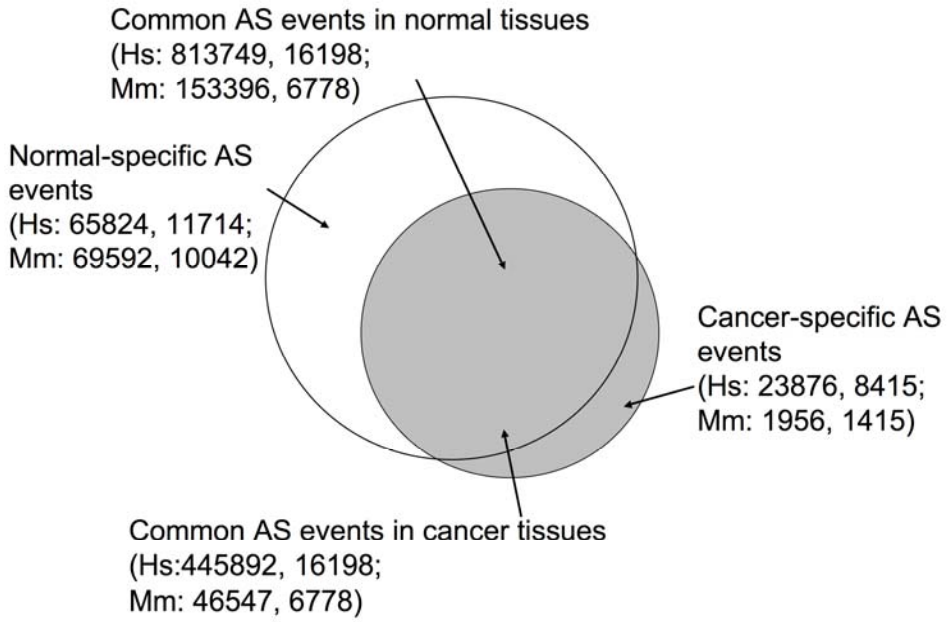


Fig. 1

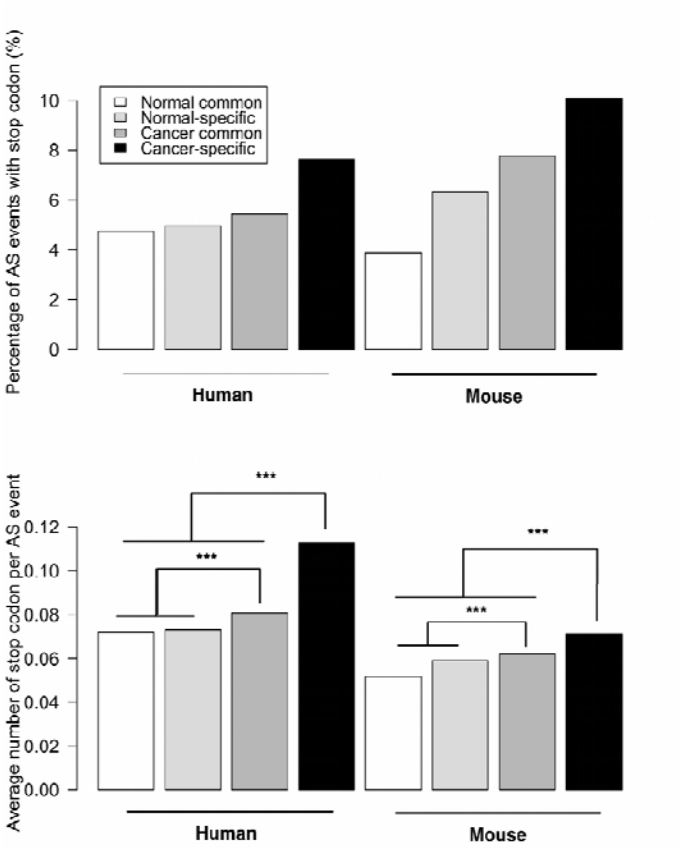


Fig. 2

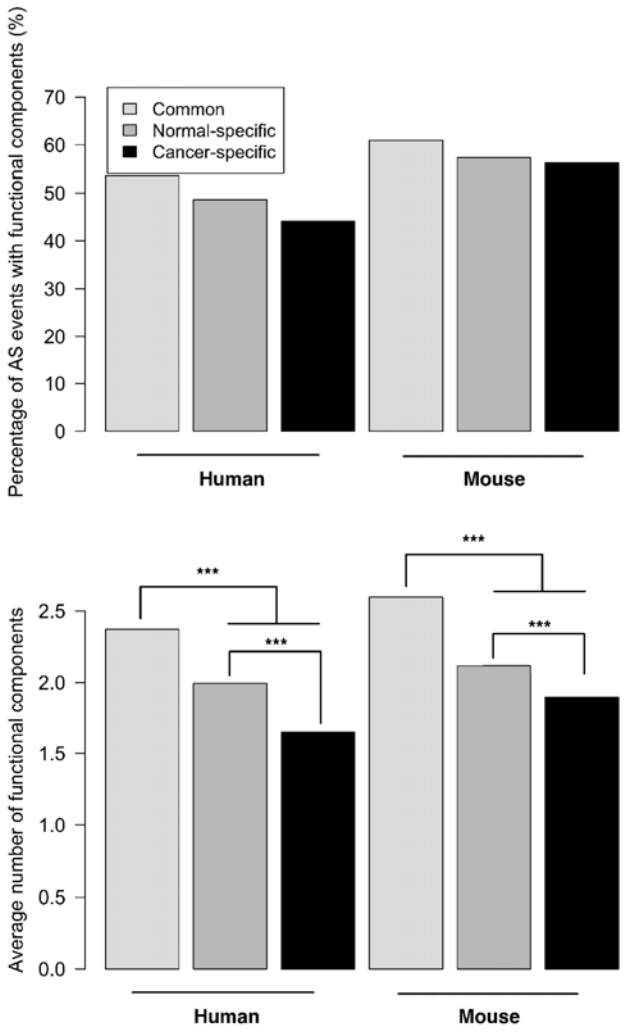


Fig. 3

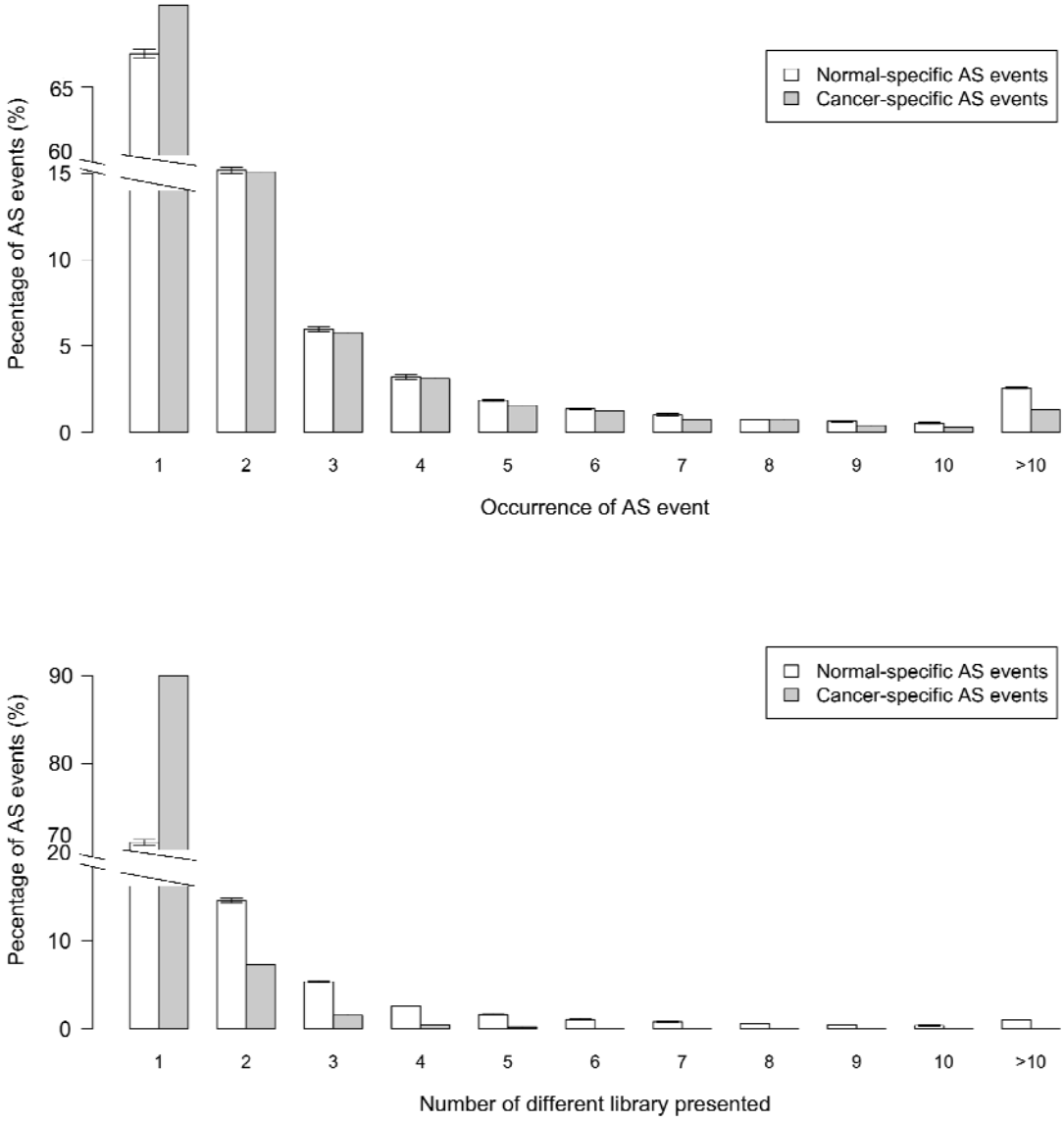


Fig. 4

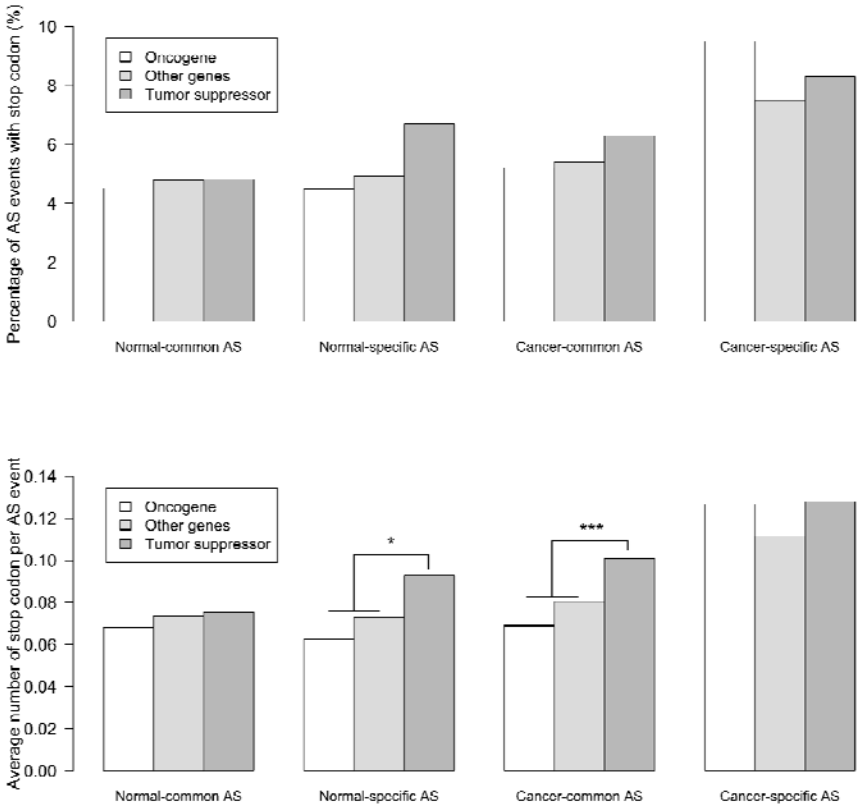


Fig. 5

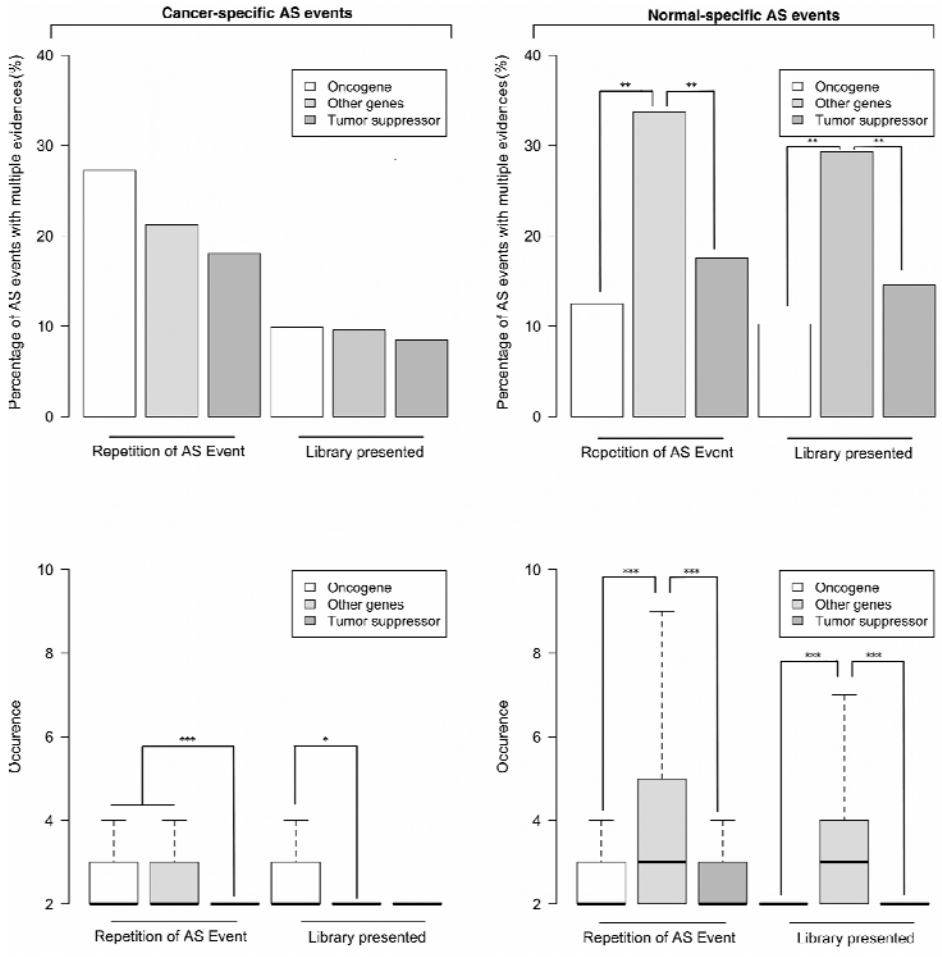


Fig. 6

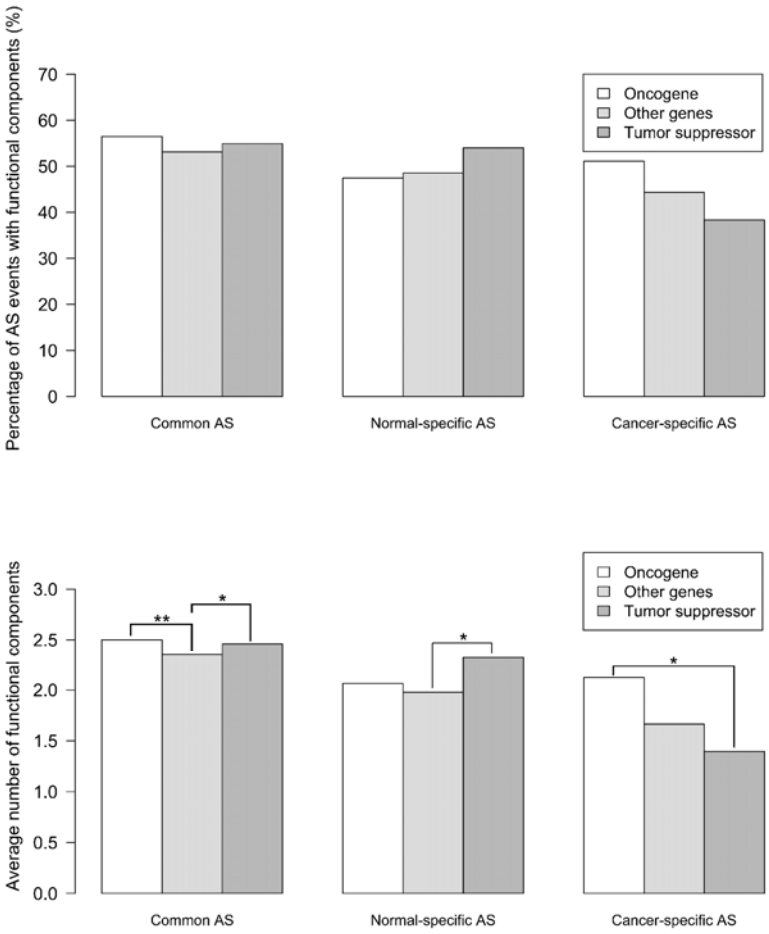


Fig. 7

TABLES

Table 1 Summary of transcripts from normal and cancer state

Species name	Disease state	Tissue type	Development stage	Library count	EST count
Human	Normal	37	7	297	1687320
	Cancer	34	5	362	920844
Mouse	Normal	29	15	164	628506
	Cancer	14	4	45	148156

Genome Biology and Evolution Advance Access published July 30, 2013 doi:10.1093/gbe/evt115

Genome Biology and Evolution

SMBE

Evidence for deep phylogenetic conservation of exonic splice-related constraints: splice-related skews at exonic ends in the brown alga *Ectocarpus* are common and resemble those seen in humans.

XianMing Wu[1], Ana Tronholm[1,2], Eva Fernández Cáceres[1], Jaime M. Tovar-Corona[1], Lu Chen[3], Araxi O. Urrutia[1] and Laurence D. Hurst[1]

1. *Department of Biology and Biochemistry, University of Bath, Bath, Somerset, UK, BA2 7AY*
2. *Present address: Department of Biological Sciences, University of Alabama, 500 Hackberry Lane, Mary Harmon Bryant Hall, Tuscaloosa, AL 35487-0345, USA*
3. *Human Genetics, Wellcome Trust Sanger Institute, Genome Campus, Hinxton CB10 1HH, UK*

Abstract

The control of RNA splicing is often modulated by exonic motifs near splice sites. Chief amongst these are exonic splice enhancers (ESE). Well described ESEs in mammals are purine rich and cause predictable skews in codon and amino acid usage towards exonic ends. Looking across species, those with relatively abundant intronic sequence are those with the more profound end of exon skews, indicative of exonization of splice site recognition. To date the only intron-rich species that have been analysed are mammals, precluding any conclusions about the likely ancestral condition. Here then we examine the patterns of codon and amino acid usage in the vicinity of exon-intron junctions in the brown algae *Ectocarpus siliculosus*, a species with abundant large introns, known SR proteins and classical splice sites. We find that amino acids and codons preferred/avoided at both 3' and 5' ends in *Ectocarpus*, of which there are many, tend, on the average, to also be preferred/avoided at the same exon ends in humans. Moreover the preferences observed at the 5' ends of exons are largely the same as those at the 3' ends, a symmetry trend only previously observed in animals. We predict putative hexameric ESEs in *Ectocarpus* and show that these are purine rich and that there are many more of these identified as functional ESEs in humans than expected by chance. These results are consistent with deep phylogenetic conservation of SR protein binding motifs. Assuming codons preferred near boundaries are "splice optimal" codons, in *Ectocarpus*, unlike *Drosophila,* splice optimal and translationally optimal codons are not mutually exclusive. The exclusivity of translationally optimal and splice optimal codon sets is thus not universal.

Introduction

While for many years patterns of biased codon usage have been typically addressed in terms of translational optimality (and fit to the tRNA pool) (Duret 2002; Sharp, et al. 2005), more recently the importance of exonic motifs involved in splicing have been seen to be relevant (Chamary and Hurst 2005; Parmley, et al. 2006; Parmley and Hurst 2007; Parmley, et al. 2007; Warnecke, et al. 2008; Willie and Majewski 2004). Chief amongst these motifs are exonic splicing enhancers (ESE) (Blencowe 2000; Cartegni, et al. 2002). At the RNA level these motifs are responsible for the binding of SR proteins to the exonic parts of the unspliced RNA, thereby enhancing splicing at the neighbouring exon-intron junction (Graveley 2000). In addition they are responsible for retaining unspliced RNA in the nucleus (Taniguchi, et al. 2007). Well described ESEs in mammals - one of the few lineages where ESEs have been experimentally confirmed (Fairbrother, et al. 2004a; Fairbrother, et al. 2002; Fairbrother, et al. 2004b; Ke, et al. 2011) - are enriched towards the ends of exons (Fairbrother, et al. 2004a), cause selective constraint at synonymous sites (Carlini and Genut 2006; Parmley, et al. 2006) and have a highly skewed nucleotide usage, being on average highly purine enriched (Fairbrother, et al. 2004a; Parmley, et al. 2007; Tanaka, et al. 1994). Well described ESEs occupy on average 30-40% of sequence near exon ends in mammals (Parmley, et al. 2006). Note that as ESEs appear to be functional up to around 70 nucleotides from an exon end (Fairbrother, et al. 2004a), exons shorter than 140bp can be considered to be all exon "end".

Owing to these three properties (high density, proximity to boundaries and skewed nucleotide content), ESEs leave a marked footprint of codon usage near exon ends of mammalian genes, with codons more commensurate with involvement in ESEs (Parmley, et al. 2007), being preferred near boundaries (Parmley and Hurst 2007). Similarly, when comparing synonymous codons, the one more employed in ESEs is relatively preferred at exon ends over the synonym (Parmley and Hurst 2007; Willie and Majewski 2004). Thus in mammals, while isochore composition is a strong driver of between-gene codon usage bias (Eyre-Walker and Hurst 2001), selection to preserve ESEs explains many of the intra-

exon trends in codon bias. Amino acids also show skews in their usage as one approaches exon-intron junctions, with trends being well predicted by nucleotide content of ESEs and the codons that contribute to any given amino acid (Parmley, et al. 2007). Indeed, comparing the usage of the two-fold blocks of leucine and arginine to their respective four-fold blocks, supports the view that these trends are both owing to nucleotide-level effects and dominantly owing to splice-related constraints (Parmley, et al. 2007). Just as knowing about ESEs makes sense of codon and amino acid trends, so too, conversely, *k*-mers that are enriched towards the ends of exons can be employed to infer nucleotide preferences of splice related motifs and to determine novel motifs (Lim, et al. 2011) (N.B. codons are in frame 3-mers).

The trends seen in mammals have a series of further properties. For example, when usage trends at the 5' and 3' ends of exons are considered separately, it appears that the trends are largely symmetrical (Lim, et al. 2011; Warnecke, et al. 2008). That is, if a codon or amino acid is highly preferred at the 5' end of exons, it is similarly highly preferred at the 3' end. The logic of this symmetry is unclear, but it may accord with a model in which SR proteins aggregate on the ends of exons within the immature RNA and this aggregate defines, by the end of the cluster, a domain where the splice junction must reside. In such a model there is no evident reason why different SR proteins should be under selection to bind 3' and 5' ends differently. However, such symmetry has to date only been observed in animals (Warnecke, et al. 2008) and not in all of them. The 5' end of exons in *Caenorhabditis* worms, for example, are not simply different in composition to the 3' ends, they show the opposite trends i.e. codons preferred at the 5' ends are avoided at the 3' ends and vice versa (antisymmetry). The 3' end trends accord with the trends seen in all other taxa, with classical purine loading. The exceptional nature of worm's 5' ends was hypothesized to reflect consequences of operonization in worm and the commensurate transplicing. The need to distinguish the 5' ends of exons from the 5' ends of genes, cut during transplicing, being suggested as the potential cause (Warnecke, et al. 2008).

More generally, the trends in codon usage at the ends of exons in mammals correlate well with those seen in other animals, for example, *Drosophila* (Warnecke and Hurst 2007). This observation is important because *Drosophila*, unlike mammals, also has evident selection for employment of "translationally optimal" codons, possibly to ensure mistranslation minimization (Akashi 1994; Drummond and Wilke 2008; Warnecke and Hurst 2010). In part the cause of the strong correlation between end of exon usage in *Drosophila* and mammals reflects the fact that the "splicing optimal" set of codons and the "translationally optimal" set of codons are two almost mutually exclusive sets of codons i.e. translationally optimal codons tend to be those avoided near exon boundaries (Warnecke and Hurst 2007). At first sight, this mutual avoidance of the two sets seen in *Drosophila* makes some sense. If the two sets were the same, in highly expressed genes SR proteins would have difficulty binding exclusively to exonic ends, as all codons would be both translationally and splice optimal. One might hence expect considerable splice disruption. Given such logic, it is worthwhile asking whether in a very distantly related species the same exclusivity rule applies.

Beyond *Drosophila*, whether the trends as observed in mammals are well conserved remains unclear as the tendency to employ SR proteins covaries with the intron density and size of introns (Warnecke, et al. 2008). This trend possibly reflects an increased tendency towards exonization of splice site recognition as introns get ever larger, small introns in a sea of large intron being the hardest to correctly splice using intronic information alone. At the other limit a species such as *Saccharomyces cerevisiae*, both shows no preference trends (Warnecke, et al. 2008), largely lacks SR proteins (Plass, et al. 2008) and has very few and small introns. The non-animal species previously analysed (such as *Arabidopsis*) have very small introns and probably don't use ESEs too commonly, although SR proteins are possibly relatively ancient within eukaryotes although poorly described outside of the animal-fungal-plant crown group (Plass, et al. 2008).

To examine whether the patterns seen in mammals might be relatively ancient, requires analysis of distant genomes with abundant and relatively large introns. To this end we selected for scrutiny the unusual genome of the brown alga *Ectocarpus siliculosus*. Brown algae share a common ancestor with the animal-fungal-plant crown group that predates the animal-fungal-plant common ancestor (Adl, et al. 2005). The genome is well sequenced and annotated (Cock, et al. 2012; Cock, et al. 2010). It is unusual in being a non-vertebrate that is rich in introns (5.1 introns per kb of exon) and those introns tend to be quite large (mean intron size = 776 bp), meaning the genome is a strong candidate for one using ESEs and SR proteins to aid splicing, with a mean CDS size to gene size ratio of 0.27, comparable to mammals (Warnecke, et al. 2008). As expected, annotation of the genome suggests it has SR proteins (Cock, et al. 2010) (see also below). The classical GT – AG rule applies in 95.3% of introns, the remainder being GC-AG introns (for sequenceLogo motifs see Figure 1; for a longer span and evidence of a classical intronic 3' polypyrimidine track see Supplementary Figure 1). Importantly, much as with humans and other intron-rich genomes, but unlike some protists and intron-poor genomes (Irimia, et al. 2007), there is not one hexameric motif that dominates intronic 5' ends (GTGAGT at 12.5% is the most common). It is thus appears an ideal candidate to ask whether the trends well resolved in humans are ancestral or animal specific. We demonstrate too that *Ectocarpus* has "translationally optimal" codons and thus ask whether these codons are never splice optimal codons.

Finally, taking advantage of what we discover to be some unusual features of the *Ectocarpus* genome, we re-examine the cryptic splice site avoidance model (Eskesen, et al. 2004). This model posits that, with introns starting GT and exons ending in G, GGT should be avoided at the 3' ends of exons (Eskesen, et al. 2004) compared to the synonym GGC. *Ectocarpus* provides an unusually "clean" test of this prediction.

METHODS

Establishing the dataset for analysis

The coding sequences (CDS) file and EMBL format exon information files for the Brown algae *Ectocarpus siliculosus* were downloaded from the database (http://bioinformatics.psb.ugent.be/genomes/view/*Ectocarpus*-siliculosus). The input CDS data was filtered to eliminate dubious sequences. We eliminated coding sequences that did not start with ATG, that did not finish with a stop codon (TAA, TAG, TGA), that had internal stop codons, that were not a multiple of three long or that contained one or more ambiguous nucleotides ("N"). In addition those where the gene sequence length does not match the sum of length of its exons as specified in the accompanying annotation files were eliminated. As we are interested in splice related constraints, gene sequences that did not contain introns were also not examined. There are in total 16579 coding sequences in the input file of which 16033 sequences qualified as suitable candidates.

Information of expression level of *Ectocarpus* genes

The EST database of *Ectocarpus* was downloaded from NCBI (http://www.ncbi.nlm.nih.gov/nucest/?term=%22Ectocarpus+siliculosus%22[porgn%3A__txid2880]). Using BLAST we identified the number of ESTs associated with each gene (identity>95%, e-value<0.01). The length corrected ESTs hit rate (EST hits divided by the length of the gene) of each gene was regarded as the relative expression level of the gene.

HMMER search for and classification of SR proteins

A SR protein reference dataset, comprised of 213 SR protein genes from different species, was established with the information from the website: http://www.bioinf.uni-leipzig.de/Leere/PRAKTIKUM/Protokolle/WS08/2/node1.html. HMMER (Eddy 1998) used to search for putative SR proteins in *Ectocarpus* genes (including those without introns) after multiple sequence alignment by MUSCLE.

To infer which, if any, of a set of cross species conserved SR proteins our candidates might belong to we performed a domain-based analysis, as previously described (Plass, et al. 2008). In brief, we examined 9 groups (families) of known SR proteins, these being: SRp20 9G8, p54 SRp86, RY1, SC35 – alias SRP1, SRm300, SRp30c-ASF, SRp40-55-75-alias SRP2, TopoI-B and Tra2. These were downloaded from: http://www.bioinf.uni-leipzig.de/Leere/PRAKTIKUM/Protokolle/WS08/2/node6.html. We aligned, using MUSCLE, the different groups of proteins separately. We then used "hmmbuild" of HMMER to make an "hmm" profile for each multiple sequence alignment. Then all profiles are collected to form a profile database. Using "hmmscan", we searched all candidate *Ectocarpus* proteins against the profile database. Finally, we determined the SR protein family that best matches each *Ectocarpus* SR candidate. To this end we considered those domains within a given *Ectocarpus* protein that are in the same order as in the reference SR protein (these being the "collinear" domains). We then sum the score of collinear domain hits for any given *Ectocarpus* protein for each reference SR protein. To choose which family a given *Ectocarpus* protein belongs to, we choose the one whose sum score of collinear domains hits is highest. Finally, we accept this classification if it the sum score of the collinear hits for a multi domain protein, or a single hit for a single domain protein, is equal to or greater than 100.

Determining trends in amino acid and codon usage

According to the information in the EMBL annotation files, we extracted every exon sequence for every qualifying gene. The trend in usage of each codon and amino acid was investigated as a function of the distance from the exon-intron boundary up to a distance of 34 codons (to accord with prior analysis (Warnecke, et al. 2008)). Importantly, the codon in direct proximity to the boundary was eliminated, but was employed to analyse splice site profiles. 5' and 3' ends were considered separately. First and last exons were excluded, leaving 95,331 exons. For each codon and amino acid under

consideration, we determined the slope on the line of proportional usage across all exons, as a function of distance from the boundary and the spearman rank correlation (rho). A negative slope on the line, or a negative rho, indicates a codon or amino acid that is preferred near exon ends, while a positive slope implies a codon or amino acid preferred at exonic cores and avoided at the ends. In prior analyses codons preferred near exon ends are well predicted by the composition of experimentally defined ESEs (Parmley and Hurst 2007).

Human exonic splice enhancer datasets

The majority of systematic attempts to define human ESEs employ computational approaches, confirmed with experimental support. Typically these approaches start with a presumption about that distribution of ESEs and look for the sequences most enriched in these trends. We analyse three such data sets. Fairbrother et al. presumed that ESEs will be enriched in exons compared with introns and more abundant in exons with weak splice sites than in those with strong splice sites (Fairbrother, et al. 2002; Fairbrother, et al. 2004b). This is the RESCUE-ESE data set. Zhang and Chasin presumed ESEs will be enriched in internal noncoding exons of protein coding genes compared to unspliced pseudo-exons and 5' untranslated regions. This is PESE dataset. Goren et al. (2006) looked for motifs that were more conserved than expected at synonymous sites and enriched, compared with background codon usage rates. This is the ESR data set. In the latter case a minority of the motifs were exonic splice inhibitors, the precise proportion being uncertain not least because ESEs can also function as exonic splice inhibitors depending on their position and context within the exon (Ke, et al. 2011). The fourth data set we consider, Ke-ESE, derives from a purely experimental approach adopted by Ke al al. (2011). They considered the effects of all possible 4096 6-mers at 5 locations in two model exons. Taking into account overlap sequences this permitted the identification of numerous ESE hexamers.

We downloaded the ESR and Ke-ESE hexamers directly from the original papers. For Ke-ESE set we selected, as the authors did, the 400 hexamers with the highest scores. RESCUE-ESE dataset was

downloaded from http://genes.mit.edu/burgelab/rescue-ese/ESE.txt and PESE original octamers from: http://www.columbia.edu/cu/biology/faculty/chasin/xz3/pese262.txt. PESE hexamers were extracted from octamers with a minimum of 7 occurrences.

Assembling a set of *Ectocarpus* putative ESEs

The attempts to infer human ESEs have, as noted above, typically specified two criteria whereby ESEs are expected to be enriched (i.e. a ***R***elative ***E***nhancer and ***S***ilencer ***C***lassification by ***U***nanimous ***E***nrichment = RESCUE method). Here we perform a similar RESCUE approach to define *Ectocarpus* ESEs. We consider that ESEs should be a) enriched at exonic ends compared with introns and b) that the usage of the ESE should increase from exon core to exon flank.

To determine the latter, for all 4096 possible hexamers we considered their relative usage in exons, in all frames, as one moves away from exon ends. We considered only those exons longer than 160 bp to ensure that enrichment at exonic ends is truly such enrichment, rather than enrichment in short exons. 5' and 3' ends were considered separately.

To consider those hexamers enriched at exon ends compared with intronic sequence, we considered exons longer than 100bp and introns longer than 100bp. We then considered the terminal 50 bp at each end of the exons and 50 bp at the end of the introns. As for statistical analysis it is important that there are the same number of introns as exons, we randomly sampled from the larger data set to equalize the size of the two.

For each hexamer we then consider its mean usage at exon ends and its mean usage at intron ends. We then calculate the difference in usage between exon end and intron end. 5' exonic ends were compared with 3' intronic ends and vice versa. For each hexamer we can then define:

$$\delta_{Observed} = exonic\ density - intronic\ density$$

To consider the significance of this we then pooled the relevant data from exons and introns, randomized them and then considered the first half of the data as being pseudo-exon and the second half pseudo-intron. Repeating this 100 times for each hexamer we define:

$$\delta_{Pseudo} = pseudo\ exonic\ density - pseudo\ intronic\ density.$$

A reasonable metric of the extent to which a given hexamer is enriched at exon ends compared with intronic ends is then:

$$Z = \frac{\delta_{Observed} - \overline{\delta_{Pseudo}}}{\sigma_{\delta_{Pseudo}}}$$

where $\overline{\delta_{Pseudo}}$ is the mean of the hexamer usage in the 100 pseudo sets and $\sigma_{\delta_{Pseudo}}$ is the standard deviation in the usage across the pseudo sets. *P* was approximated by extrapolation from *Z* under an assumption of normality.

To generate a set of ESEs we then consider those hexamers enriched in exon end compared with intron ($Z>0$), preferred near exon ends compared with core (rho<0, slope <0) and combine *P* values from the two approaches using Fisher's method. We then consider those hexamers with a combined *P* <0.05/4096 as putative ESEs.

CAI calculation, identification of "optimal" codons and the relationship with gene expression

A dataset containing 43 ribosomal proteins was established and employed as a reference "highly expressed" gene class. The codon usage in this set was analysed using CodonW. We employed this reference dataset, and the reference codon usage table from Codon Usage Database (www.kazusa.or.jp/codon/countcodon.html) to determine codon adaptation index (CAI) scores for all genes. To this end we downloaded the local version of CAIcal, a CAI calculation program, from

http://genomes.urv.es/CAIcal/ and calculated the CAI values of all genes, excepting the ribosomal genes which had been used for establishing reference dataset. The validity of the CAI index was examined by considering the relationship between expression level and CAI, again excluding the ribosomal genes. For any given synonymous block the codon with the highest codon adaptation index was considered to be the "optimal" codon. tRNA copy numbers were obtained from http://plantrna.ibmp.cnrs.fr/.

Alternative splicing event calculation

Alternative splicing events in human and *Ectocarpus* genes were identified from 8315122 and 67082 EST sequences respectively, downloaded from dbEST database (Boguski, et al. 1993), using methods previously outlined (Chen, et al. 2011). In brief, individual ESTs were matched to individual genes by aligning them to the genome sequence using GMAP (Wu and Watanabe 2005). Exon templates were then inferred from EST alignment coordinates. Alternative splicing events were identified by comparing alignment coordinates for each EST against the exon template.

Comparable alternative splicing even counts correcting for EST coverage were obtained using a transcript normalization protocol as described previously (Kim, et al. 2007), where alternative splicing events per gene are calculated as the average number of alternative splicing events identified in 100 random samples of 10 ESTs.

RESULTS

***Ectocarpus* has multiple SR proteins**

Before asking whether binding of SR proteins to ESEs leaves a footprint of biased codon and amino acid usage in proximity to intron exon boundaries, we first established the profile of SR proteins within the genome. To search for candidates we did a HMMER search, training the HMMER on an established collection of SR proteins. In total, we identified 54 putative SR proteins, including three previously annotated as SR proteins (Esi0638_0002, Esi_0327_0029, Esi0164_0021) (supplementary result 1). In the original build of the genome a further putative SR protein was identified (Esi0089_0034; annotated as "splicing factor, arginine/serine-rich 2, RNAp interacting protein, putative"). This wasn't identified by HMMER. While many of the extra hits are unlikely to be SR protein (e.g Eukaryotic translation initiation factor 3 subunit a), several more have suggestive RNA binding functions. Most of the extra hits are not annotated.

To clarify just which SR proteins the HMMER search might have revealed we performed an additional domain based analysis, comparing our proteins to nine SR proteins that are relatively well conserved through plants, animals and fungi (Plass, et al. 2008). We find robust evidence for 18 of the *Ectocarpus* genes as being members of one of 5 SR protein families (Figure 2, Supplementary Table 1). This includes the three previously annotated SR proteins (Supplementary Table 1). We conclude that *Ectocarpus* has a good number of SR proteins, but probably not the full set described in humans (Long and Caceres 2009).

A high proportion of amino acids and codons show preference/avoidance trends

To determine which, and how many, codons and amino acids show significantly skewed usage in proximity to exon-intron junctions, we considered the relative usage of all codons and amino acids as a function of distance from an exon-intron junction, ignoring the codon in immediate proximity to the junction. We examine 3' and 5' ends separately. Any codon or amino acid preferred near a boundary will have a negative slope and a negative spearman rank correlation (rho) between its relative usage and

the distance from the boundary. A positive slope/rho score indicates avoidance near a boundary relative to usage more core to exons. The slope/rho values we consider to be measures of the preference/avoidance trends.

Most codons and amino acids show significant preference/avoidance directions near exon-intron boundaries (supplement tables 2 & 3, supplementary figures 2 & 3). Before Bonferonni correction 86% of codons and 96% of amino acids show significant trends ($P<0.05$) at the 5' exonic ends and 88% of codons and 91% of amino acids show significant trends ($P<0.05$) at the 3' exonic ends. After correction, these numbers drop to 68% of codons and 83% of amino acids show significant trends at the 5' exonic ends and 69% of codons and 83% of amino acids show significant trends at the 3' exonic ends. Overall, considering all codons with at least one synonym, 66% of comparisons show significant trends after Bonferonni multitest correction and 83% of amino acids analyses show significant trends.

These figures compare strikingly with what has been seen before. *A priori* we expect that species with relatively small exons sitting in an intron-rich sea will be those that will be under selection for adding exonic splice information to bolster the intronic signals. This should in turn be reflected in more codons and more amino acids showing skewed usage near boundaries. This supposition is generally supported by the finding that species in which the ratio of the mature CDS size to gene size is small are those in which a greater proportion of amino acids or codons show significant skews towards exon ends (Figure 3). When we consider our new data in this light, while *Ectocarpus* certainly has a low CDS to gene ratio, the proportion of amino acids showing a skew remains unusually high (Figure 3). We note that this cannot be an artifact of sample sizes (the higher the sample size the more likely significant skews will be seen even if trends are weak) as humans and mice have fewer significant trends but more exons analysed.

Preference/avoidance trends at 3' and 5' exon ends are similar

Is the pattern of symmetry seen in most animals, but not so far reported outside of animals, also seen in *Ectocarpus*? To address this we considered for each amino acid and each codon, the trend in its usage approaching the 5' and 3' ends of exons. The slope on this line and the spearman rho values were considered. We then consider the correlation between the figures when comparing 5' and 3' ends. We find that overall exons tend to be symmetric, with a strong correlation both between the slopes and the rho values for codons (Figure 4: from binomial test with success as preservation of the sign of the slope, P=0.004; from spearman rank correlation on slopes, P=0.001) and amino acids (Figure 5: from binomial test with success as preservation of the sign of the slope, P=0.0026, from spearman rank correlation on slopes, P=0.002).

However, closer scrutiny indicates a further nuance, namely that C and G ending codon usage can be antisymmetrical (Figure 4). CTC, GTC and ACC are all highly disfavoured at exonic 5' ends but highly preferred at 3' ends, while GTG, GAG, ACG, CGG and CTG all show the opposite pattern. Generally at four fold-degenerates sites, C and G show somewhat antisymmetrical patterns, with C avoided at 5' ends while G, although highly abundant, is avoided at 3' ends, meaning as one approaches the boundary its usage declines (Figure 6,a, b). However, these patterns are not simple enough to be explained solely in terms of C or G content, as many C ending codons, including CCC, are symmetrical. Nonetheless, we conclude that the symmetry rules are not universally respected.

Usage trends near exon junctions in *Ectocarpus* resemble those seen in humans.

To consider whether the trends seen in humans are also seen in a very distant species with large intronic content such as *Ectocarpus*, we considered how the rho values for each amino acid, 5' and 3' resembled those seen in humans. We can then ask two questions. First, do the trends overall correlate between the two species? Second, what proportion of comparisons has a conserved direction of preference/avoidance between the two species?

To this end we consider the correlations between the trends (rho values) seen in the two species, considering 5' and 3' ends separately. Remarkably, despite over 1 billion years of divergence, we find that at both 5' and 3' exonic ends the trends correlate well (5' end: rho=0.68, *P*=0.0005; 3' end: rho=0.53, *P*=0.01: Figure 7; Table 1). At 3' ends only 4 of the 23 codon blocks (we treat 2 fold and 4 fold blocks differently here) do not have a conserved preference trend (binomial test, *P*=0.002). At 5' ends while the trends are correlated, 7 codons blocks have different trends, rendering the result non-significant (binomial test, *P*=0.09). However, when we restrict analysis to those amino acids showing significant trends (before Bonferonni correction), we observe that at both 5' and 3' ends the proportion conserved is significant (Table 1). Moreover, we note that of the 4 amino acids that are antisymmetrical in *Ectocarpus* (E, R, V and T), three (E, V and T) are also antisymmetrical in humans (Supplementary Fig 4).

We can perform a similar analysis using codons (excluding ATG and TGG owing to lack of synonyms). Again we find strong concordance between trends seen in humans and those seen in *Ectocarpus* (5' end: rho=0.50, $P=5.17 \times 10^{-5}$; 3' end rho=0.58, $P=1.7 \times 10^{-6}$, Figure 7; Table 1). The conservation patterns mirrors what we see at the amino acid level. At the 3' ends we see a strong correlation and only 12 show reverse trends (binomial test, $P=5.13 \times 10^{-6}$). At the 5' ends the effects are more modest. A significant correlation is observed but of 59 codons, 23 show reverse trends at 5' ends (*P*=0.12). As above, restricting analysis to only those codons showing significant trends, at both 5' and 3' ends more than expected show conservation of direction (Table 1). Nucleotide usage at four fold degenerate sites is also comparable between *Ectocarpus* (Figure 5a, b) and humans (Figure 5, c, d), although in *Ectocarpus* the C and T preferences at the 3' end are more similar than seen in humans. Overall, these results suggest a deep phylogenetic conservation of splice associated trends in amino acid composition as one approaches exonic ends, most especially at the 3' ends.

Ectocarpus has low rates of alternative splicing

The similarity that we see between humans and *Ectocarpus* in terms of which codons and amino acids are preferred and avoided near boundaries suggests that the selection on the nucleotide usage in DNA or RNAs at exon ends is for similar reasons. Why then does *Ectocarpus* have so many more amino acids and codons showing significant trends (Figure 2)? The metric we employ is by no means perfect as it is sensitive to sample sizes (number of exons examined). However, high numbers of trends seen for *Ectocarpus* compared with mouse/human cannot be an artifact of sample sizes, as the sample sizes in vertebrates (in terms of number of exon ends) is higher than that in *Ectocarpus*.

One possibility to explain the large number of skews in *Ectocarpus* is that alternative splicing might be relatively rare in *Ectocarpus*. The consequence of this would be that most exons are consistently under strong selection to be spliced correctly. By contrast, if in humans many exons are splicing errors (Zhang, et al. 2009), then we would not expect strong selection to preserve ESEs in all exons. The uniformity of splice sites in *Ectocarpus* (Figure 1) would be consistent with the hypothesis that most exons are under selection to be properly spliced.

Preliminary data suggests that alternative splicing is indeed rare in *Ectocarpus*. A detailed examination of splice forms has been performed on one gene family, the cytosolic glutathione transferases. While eleven genes were identified only one had an alternative transcript (Franco, et al. 2008). While this is very much lower then the rate seen in humans, where nearly all intron bearing genes have at least two isoforms (Pan, et al. 2008), this difference might reflect, at least in part, differences in the depth of study (Brett, et al. 2002).

In order to compare alternative splicing levels in both human and *Ectocarpus* allowing for depth of EST sequencing, we performed two tests. First, we measured alternative splicing levels in genes from both species after transcript number normalisation. For this, alternative splicing per gene was measured as

the average number of alternative splicing events detected in 1000 random samples of 10 ESTs. We obtained this comparable index of alternative splicing for 8772 human genes and 69 *Ectocarpus* genes. We found that while *Ectocarpus* genes had an average of 0.41 events per gene (median of 0), human genes had an average of 5.35 events per gene (median of 4.55). This difference is highly significant (t-test, $P=2.15 \times 10^{-68}$). Second, we compared the average number of alternative splicing events detected when genes are grouped according to the number of ESTs aligning to them. When genes were divided according to their average number of aligned ESTs, the average number of alternative splicing events per gene was considerably higher for human compared to *Ectocarpus* at all 9 EST per gene counts ($P=0.004$ from binomial test: N= 4861 human; 326 *Ectocarpus,* Figure 9). We conclude that in *Ectocarpus* alternative splicing is rare compared to that seen in humans. While this is consistent with the possibility that alternative transcription rates might impact on the net skew in nucleotide usage, this hypothesis requires considerable further cross-taxon analysis.

Ectocarpus putative exonic splice enhancers resemble those seen in humans

Above we compared human and *Ectocarpus* exonic ends as regards trends in codon usage. The trends seen in mammals reflect the nucleotide content of ESEs (Parmley and Hurst 2007). This is to be largely expected as ESEs are hexamers that tend to be enriched at exon ends in any frame and codons are 3-mers in frame and hence are likely to be non-independent of ESE imposed trends. ESEs are however not described in *Ectocarpus* so we cannot perform the same analysis. We can however attempt to determine which hexamers might function as ESEs and compare this set of candidates with those identified in humans.

To this end we asked of *Ectocarpus* a) which hexamers are enriched at exonic ends compared with intronic sequence and b) which hexamers are used at exon ends more than in exon centres (i.e. have a negative slope of proportional usage against distance from exonic end). We then consider as candidate

ESEs the hexamers most enriched on both axes. Note that this method is far from perfect in so much as we also identify causes of skew in codon usage probably not related to ESEs but rather to avoidance of cryptic splice sites (see below).

We identify 904 3' ESEs and 919 5' ESEs (Supplementary Table 4). The 5' and 3' hexamers are different from each other. We observed 189 in common but by chance we would expect 203. Moreover, while the 5' hexamers are, like classically described SR protein binding ESEs, highly purine enriched (A+G = 64.4%), the 3' set are if anything pyrimidine enriched (A+G=42%), this being consistent with the C enrichment at 4 fold sites at the 3' exonic ends (Figure 6). The set of 189 ESEs that are common to 5' and 3' ESEs are purine rich (A+G= 59.3%).

There are four high-throughput data sets attempting to identify ESEs in humans (see Methods). Unfortunately these four have remarkably few hexamers in common – just 10 of over 900 putative hexamers are found in all four data sets. We consider those hexamers found in at least 3 of the four data sets as being a robust set of human ESEs (N=54). For both our 3' and 5' set of hexamers we find considerably more overlap than expected under a null model in which the human set of ESEs and the *Ectocarpus* set are assumed to be independent. For the 5' ESEs we expect 12 hexamers in common but observe 39, more than 8 standard deviations more than expected by chance ($P<<0.0001$). For the 3' end ESEs the effect is more modest but still highly significant: we observe 26 in common between the two sets, where less than 12 are expected by chance, this being nearly 5 standard deviations away ($P<0.0001$). Of the 189 hexamers that are in common at 5' and 3' ends, 18 are also in the set of 54 human ESEs while less than 3 are expected under a random null. This deviation is almost 10 standard deviations from expectations ($P<<0.0001$). All of these degrees of concordance between *Ectocarpus* and human are considerably greater in magnitude than the concordance witnessed between some of the initial four human data sets. We conclude that despite the unusual base composition of 3' ESEs in *Ectocarpus*, there is a significant resemblance between human ESEs and *Ectocarpus* ESEs. The trends,

especially those seen at the 5' ends, is consistent with a deep and strong phylogenetic conservation of SR protein binding preferences.

Translationally optimal and splice optimal codons are not mutually exclusive in *Ectocarpus*

In *Drosophila* the set of codons enriched near exon ends accords with those commensurate with ESEs, and correlate well with the trends seen in mammals (Warnecke and Hurst 2007). These "splice optimal" codons are very different from the "translationally optimal" codons, with just one codon being in both sets (Warnecke and Hurst 2007). Is this mutual exclusivity also seen in *Ectocarpus*? To address this we first must ask whether *Ectocarpus* is like *Drosophila* in having a translationally optimal class of codons. To this end we first examined codon usage in the ribosomal proteins, these being the most highly expressed genes. Given the difference in the codon usage in the ribosomal genes and the codon usage in the genome as a whole, we could then ascribe each gene a CAI score. We then ask whether, excluding the ribosomal protein training set, the more highly expressed genes show higher CAI. There is a weak but significant correlation between CAI and expression level (Pearson correlation, r=0.084; P=2.6 x 10^{-13}, Supplementary Figure 5).

In addition, we compared the optimal codons, as defined by over usage in ribosomal proteins, for each synonymous set with the tRNA copy numbers (assuming these to be a rough guide to tRNA levels) and asked if the optimal codon within each block was also the one with the most abundant tRNA. In 12 of 18 synonymous blocks this was the case (Supplementary Table 5). By randomizations, involving extracting at random 2 codons from each synonymous block, we asked how often we expected by chance to see 12 of 18 matching, given the structure of the genetic code. In 100,000 simulations in less than 1000 incidences did we observe 12 or more matches. We conclude that the optimal codons tend to be those matching the more abundant tRNAs (P<0.001). *Ectocarpus* is, in this regard, more like flies than mammals, and is under translational selection.

Given the above result, we can now address whether the "translationally optimal" codons might be different from the splice optimal codons. To define splice optimal codons we consider all those preferred near exon boundaries (at both 5' and 3' ends) that are significantly skewed after Bonferonni correction ($P<0.05/118$ at both ends and rho <0) (Supplementary Table 5). This defines 16 codons, although some of these are from the same codon block. Indeed, of 18 amino acids with more than one synonym, 10 amino acids have no splice preferred codons. In the remaining cases, three amino acids have all their codons as splice optimal (F, I, K, Y). In three (H,L,R) of the remaining four informative cases the translationally optimal codon, defined by reference to usage in ribosomal proteins, is not a splice optimal codon, but in K it is. To examine the significance of this we considered a simulation in which we define for each of the four codon blocks the number of splice optimal codons and randomly sampled that number out of the number of codons in the block. We then ask how often the pseudo-splice set of codons and the pseudo translationally optimal codon matches. We then consider how often we see 1 or fewer matches. We find that we expect this to happen about 41.6% of the time, thus there is no evidence that splice optimal and translational optimal codons are under selection to differ.

We can be less stringent and define a splice optimal codon as any codon showing preference towards any exon end (not both 5' and 3' ends) after Bonferonni correction ($p<0.05/118$). This gives 34 splice optimal codons (Supplementary Table 5). There are here only 4 potentially informative synonymous codon blocks in which some but not all of the codons are splice optimal. As regards translational optimality defined in terms of usage in ribosomal proteins, N and T have splice optimal codons that are translational optimal ones, while Q and R have the opposite. Again we see no significant evidence that splice optimal and translational optimality are divergent (from simulation: $P=0.65$).

Additionally, for each codon block we can ask which codon is the most splice preferred. This we define as the codon with the most significantly negative slope using both 5' and 3' analyses. If no

codon has a significantly negative slope then we consider the one with the most negative slope to be the splice preferred codon. We find that in nine of 18 incidences the splice preferred codon is also the translationally optimal codon. We reject the hypothesis that splice optimal codons tend not to be translationally optimal codons (by simulation: P=0.91).

Evidence of cryptic splice site avoidance

The nucleotide composition at 3' exonic ends, allows us to provide an unusually "clean" test of the cryptic splice site avoidance model (Eskesen, et al. 2004). Given that introns start GT and end AG, to avoid cryptic splice sites, it is argued (Eskesen, et al. 2004) that AG residues should be avoided near 5' ends of exons and GT should be avoided at the 3' ends. One difficulty with any such analysis in mammals, however, is that, nucleotide usage in ESEs trends to go in the same direction as predictions from the cryptic splice avoidance model (Chamary and Hurst 2005). *Ectocarpus* provides an opportunity to test the cryptic splice model as at the 3' ends both T and C are weakly preferred and show very similar relative trends (Fig 6b). As exons tend to end G in the majority of incidences (Figure 1), the cryptic splice avoidance model thus predicts that at 3' exon ends GGT should be avoided compared to GGC (a cryptic splice could occur between the two G residues in GGT), but [A|C|T]GT need not be avoided compared with [A|C|T]GC. Precise expectations for [A|C|T]GC and [A|C|T]GT are not however clear, their relative usage potentially reflecting background nucleotide trends. Given this we ask solely whether GGT/GGC behaves differently from [A|C|T]GC and [A|C|T]GT, with the latter three consistent in their behavior. We observe just this, with profound avoidance of GGT compared to GGC, but preference for [A|C|T]GT, compared with [A|C|T]GC, near boundaries at the 3' ends of exons (Figure 10). Note too that both GGN and CGN are four-fold degenerate codons so this comparison is especially well controlled. At the 5' ends of exons the pattern is reversed with GGT being preferred over GGC, which is to be expected given the overall nucleotide composition, C being

strongly avoided 5' and T being weakly preferred. In sum, the preference of GGC over GGT at exonic 3' end is consistent with the cryptic splice site model.

At the exonic 5' end as introns end AG and exons commonly start with a G (Figure 1), the cryptic splice model predicts that AGG should be avoided compared with AGA. This is observed (Supplementary Figure 6). This test is not a strong one however as, while exons regularly end G, the preference to start with a G is weaker.

While the cryptic splice model makes good sense of the preference for GGC over GGT at 3' ends, most other trends cannot be explained in terms of splice avoidance. For example, preference for [A|C]GT over [A|C]GC most probably reflects processes acting more generally. We presume, as typically done (Lim, et al. 2011), that most of the trends observed reflect the nucleotide content of splice motifs, such as ESEs.

Discussion

The analysis of the *Ectocarpus* genome has provided the first insight into the splice related forces operating in a very distant relative of vertebrates in a species with intronic content comparable to that of vertebrates. The extent to which the trends observed in vertebrates accord with those seen in *Ectocarpus* are striking given the vast evolutionary distance between the two groups. It seems, therefore, parsimonious to presume that this reflects splice related constraints, most probably the conservation of the binding motifs of SR proteins, not least because trends seen in humans accord well with those expected given the nucleotide composition of ESEs (Parmley and Hurst 2007; Parmley, et al. 2007). Moreover, that *k*-mer enrichment in the vicinity of exon boundaries is a successful method to identify new splicing motifs, supports the supposition that the codon trends that we observe reflect splice related motifs (Lim, et al. 2011). The correspondence between our predicted set of *Ectocarpus* ESE hexamers and human ESEs is noteable. It is to be expected that, just as vertebrate ESEs can function

in fungi (Webb 2005), so too they might function in brown algae. Nonetheless, given our results regarding the cryptic splice site avoidance model, we see no reason to suppose that ESE enrichment is the sole cause of all the trends that we observe.

Why *Ectocarpus* has so many codons and amino acids showing strong preference avoidance trends (and also so many putative ESEs) is unclear. The possibility that alternative splicing is rare in *Ectocarpus*, hence resulting in selection on most exons most of the time for correct splicing, is consistent the data but requires further scrutiny. As regards the trends seen at the amino acid level, an alternative possibility to splice related selection is that we are detecting preference for one amino acid above another, owing to the hypothesised tendency of protein modules to reside in individual exons, as conjectured by the introns-early hypothesis (Gilbert, et al. 1986). Aside from the fact that the one-module one-exon hypothesis is probably untenable (Logsdon 1998; Stoltzfus, et al. 1994), this possibility is rejected in humans, not least because of the 6-fold degenerate amino acids, two (L and R) show opposite trends within the two-fold and four-fold degenerate blocks, these trends being well predicted by involvement in ESEs (Parmley, et al. 2007). Similarly, for the two-fold block of arginine (R) in *Ectoparpus*, at the 5' exon ends both codons are avoided, while the three of the four codons of the four-fold block are preferred. Within the four-fold blocks of both valine and threonine there are both codons that are significantly avoided and significantly preferred. Similarly, within the two fold degenerate block of arginine and the two-fold degenerate glutamic acid at 3' exon ends one of the two is significantly avoided and one is significantly preferred. These differing trends within synonymous codon blocks support the hypothesis that, at least in part, the trends that we observed are owing to nucleotide level, not protein level, effects. Nonetheless, most pairs of codons in two-fold degenerate blocks have preference trends in the same direction. This could reflect either some relationship to protein structure or similar splice related selection (e.g. ESE involvement) owing to the first two bases in the two-fold degenerate codons being identical in the synonyms.

Unlike *Drosophila*, *Ectocarpus*, while having evidence of being under translational selection, shows no evidence of selection to make distinct splice optimal and translationally optimal codons. Why might the two genomes differ? One possibility is that selection for translational optimality is that bit stronger in *Drosophila*. Indeed in *Ectocarpus* the correlation between CAI and expression level is rather weak. Were this weakness real (as opposed to an artifact of limited and noisy expression data) then selection to force divergence between translationally optimal and splice optimal codons may also be weak. Another possibility is that the observation in *Drosophila*, while seemingly having an attractive explanation, is an accidental consequence of selection on translational optimality and splice optimality happening to go in opposite directions. Until it is better understood why certain codons end up being translationally optimal this issue will be hard to resolve. Nonetheless we can now provide an exemplar where translational optimality has not obviously selected on a set of codons distinct from the splice optimal set.

Acknowledgements

We thank Simon Dittami for advice on expression resources. JMTC is funded by a CONACyT scholarship, AOU is a Royal Society Dorothy Hodgkin Research Fellow, LDH is a Royal Society Wolfsom Research Merit Award Holder. XW is funded by the University of Bath. EFC is supported by the Erasmus program.

References

Adl SM, et al. 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J. Eukaryot. Microbiol. 52: 399-451.

Akashi H 1994. Synonymous codon usage in *Drosophila* melanogaster: Natural selection and translational accuracy. Genetics 136: 927-935.

Blencowe BJ 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem. Sci. 25: 106-110.

Boguski MS, Lowe TM, Tolstoshev CM 1993. dbEST--database for "expressed sequence tags". Nat. Genet. 4: 332-333.

Brett D, Pospisil H, Valcarcel J, Reich J, Bork P 2002. Alternative splicing and genome complexity. Nat. Genet. 30: 29-30.

Carlini DB, Genut JE 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. J. Mol. Evol. 62: 89-98.
Cartegni L, Chew SL, Krainer AR 2002. Listening to silence and understanding nonsense: Exonic mutations that affect splicing. Nat. Rev. Genet. 3: 285-298.
Chamary JV, Hurst LD 2005. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? Trends Genet. 21: 256-259.
Chen L, Tovar-Corona JM, Urrutia AO 2011. Increased levels of noisy splicing in cancers, but not for oncogene-derived transcripts. Hum. Mol. Genet. 20: 4422-4429.
Cock JM, et al. 2012. The *Ectocarpus* Genome and Brown Algal Genomics The Ectocarpus Genome Consortium. In: Piganeau G, editor. Genomic Insights into the Biology of Algae. Amsterdam: Elsevier Ltd. p. 141-184.
Cock JM, et al. 2010. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. Nature 465: 617-621.
Drummond DA, Wilke CO 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341-352.
Duret L 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. 12: 640-649.
Eddy SR 1998. Profile hidden Markov models. Bioinformatics 14: 755-763.
Eskesen ST, Eskesen FN, Ruvinsky A 2004. Natural selection affects frequencies of AG and GT dinucleotides at the 5 ' and 3 ' ends of exons. Genetics 167: 543-550.
Eyre-Walker A, Hurst LD 2001. The evolution of isochores. Nat. Rev. Genet. 2: 549-555.
Fairbrother WG, Holste D, Burge CB, Sharp PA 2004a. Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol. 2: E268.
Fairbrother WG, Yeh RF, Sharp PA, Burge CB 2002. Predictive identification of exonic splicing enhancers in human genes. Science 297: 1007-1013.
Fairbrother WG, et al. 2004b. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. Nucleic Acids Res. 32: W187-190.
Franco P-Od, Rousvoal S, Tonon T, Boyen C 2008. Whole genome survey of the glutathione transferase family in the brown algal model *Ectocarpus siliculosus*. Mar. Genomics 1: 135-148.
Gilbert W, Marchionni M, McKnight G 1986. On the antiquity of introns. Cell 46: 151-154.
Goren A, et al. 2006. Comparative analysis identifies exonic splicing regulatory sequences - The complex definition of enhancers and silencers. Mol. Cell 22: 769-781.
Graveley BR 2000. Sorting out the complexity of SR protein functions. RNA 6: 1197-1211.
Irimia M, Penny D, Roy SW 2007. Coevolution of genomic intron number and splice sites. Trends Genet. 23: 321-325.
Ke S, et al. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. 21: 1360-1374.
Kim E, Magen A, Ast G 2007. Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. 35: 125-131.
Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG 2011. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. Proc Natl Acad Sci U S A 108: 11093-11098.
Logsdon JM 1998. The recent origins of spliceosomal introns revisited. Curr. Opin. Genet. Dev. 8: 637-648.
Long JC, Caceres JF 2009. The SR protein family of splicing factors: master regulators of gene expression. Biochem. J. 417: 15-27.
Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. 40: 1413-1415.
Parmley JL, Chamary JV, Hurst LD 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol. Biol. Evol. 23: 301-309.

Parmley JL, Hurst LD 2007. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. Mol. Biol. Evol. 24: 1600-1603.
Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD 2007. Splicing and the evolution of proteins in mammals. PLoS Biol. 5: 343-353.
Plass M, Agirre E, Reyes D, Camara F, Eyras E 2008. Co-evolution of the branch site and SR proteins in eukaryotes. Trends Genet. 24: 590-594.
Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE 2005. Variation in the strength of selected codon usage bias among bacteria. Nucl. Acids Res. 33: 1141-1153.
Stoltzfus A, Spencer DF, Zuker M, Logsdon JM, Doolittle WF 1994. Testing the exon theory of genes - the evidence from protein-structure. Science 265: 202-207.
Tanaka K, Watakabe A, Shimura Y 1994. Polypurine sequences within a downstream exon function as a splicing enhancer. Mol. Cell. Biol. 14: 1347-1354.
Taniguchi I, Masuyama K, Ohno M 2007. Role of purine-rich exonic splicing enhancers in nuclear retention of pre-mRNAs. Proc. Natl. Acad. Sci. USA 104: 13684-13689.
Warnecke T, Hurst LD 2007. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. Mol. Biol. Evol. 24: 2755-2762.
Warnecke T, Hurst LD 2010. GroEL dependency affects codon usage-support for a critical role of misfolding in gene evolution. Mol. Syst. Biol. 6: 340.
Warnecke T, Parmley JL, Hurst LD 2008. Finding exonic islands in a sea of non-coding sequence: splicing related constraints on protein composition and evolution are common in intron-rich genomes. Genome Biol. 9: r29.
Webb CJ 2005. Exonic splicing enhancers in fission yeast: functional conservation demonstrates an early evolutionary origin. Genes Dev. 19: 242-254.
Willie E, Majewski J 2004. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. Trends Genet. 20: 534-538.
Wu TD, Watanabe CK 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21: 1859-1875.
Zhang Z, et al. 2009. Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay. BMC Biol. 7: 23.

Tables

Table 1. Conservation of trends between humans and *Ectocarpus*. The trends in usage of all codons and amino acids at 3' and 5' ends of exons were compared between human and *Ectocarpus*. Two reporting statistics were considered. First we ask using a binomial test whether the proportion of observations changing direction of trend is different from expected under a null where trends are free to evolve. Second we consider a Spearman Rank correlation test. As the analysis can be biased by considering trends that are very marginal and non-significant we perform a second analysis where only significant trends (P<0.05 before Bonferonni correction in both species) are employed.

		Binomial test		Spearman's Rank Correlation	
		Changed direction	*P*	*Rho*	*P*
All observations	5' AA	7 from 23	0.093	0.6779	0.0005
	3' AA	4 from 23	0.0026	0.5316	0.0100
	5' codon	23 from 59	0.1175	0.5017	5.17E-05
	3' codon	12 from 59	5.13E-06	0.5773	1.70E-06
Significant observations	5' AA	3 from 16	0.021	0.7559	0.0011
	3' AA	3 from 17	0.013	0.5000	0.0430
	5' codon	11 from 43	0.0019	0.5694	6.75E-05
	3' codon	5 from 38	4.26E-06	0.5938	8.51E-05

Figure 1. Splice site composition in *Ectocarpus*.

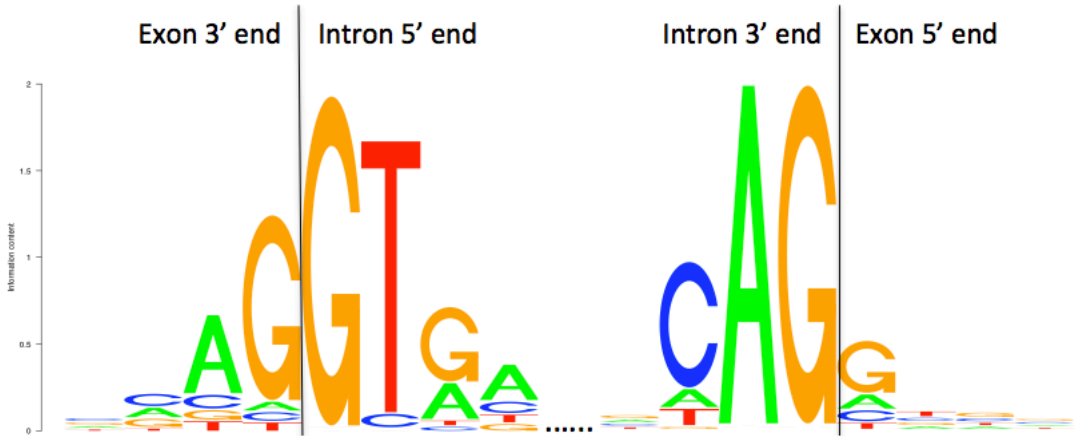


Figure 2: Presence or absence of SR and SR related proteins in *Ectocarpus* compared with other taxa. Presence of at least one member of a gene family in a given species is indicated in dark blue or absence in light yellow. Data for all species bar *Ectocarpus* from Plass et al. (2008).

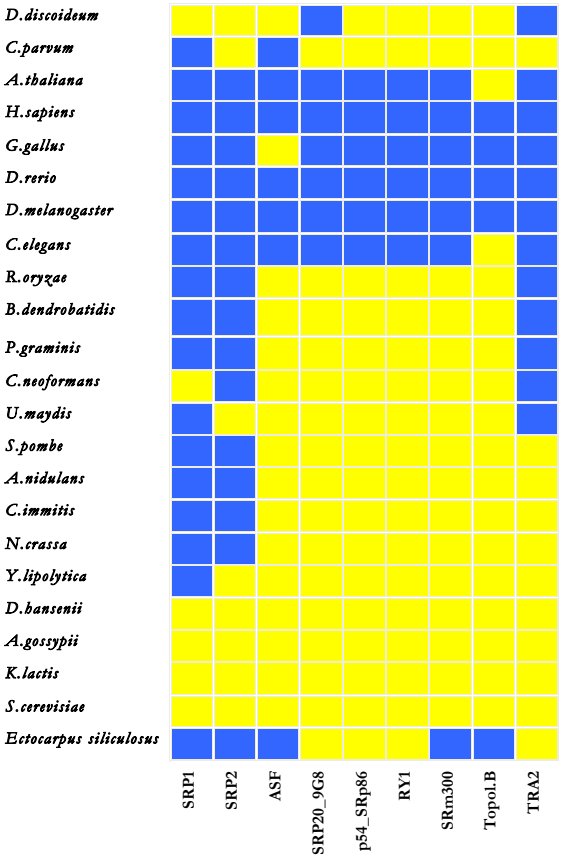


Figure 3. The proportion of amino acids showing significant preference/avoidance trends after Bonferonni correction as a function of the average ratio of mature CDS to gene length across multiple species.

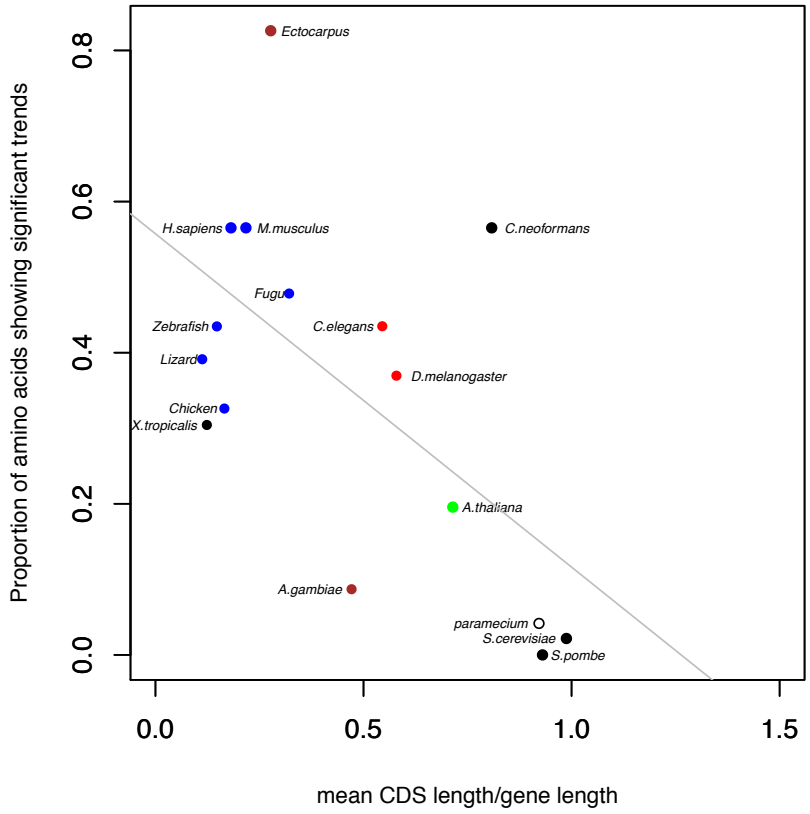


Figure 4. Examination of symmetry of preference/avoidance trends for codons

Both 5' and 3' exon ends we considered both **a.** the slope on the line of relative usage versus distance from the boundary (rho=0.40, *P*=0.0014) and **b.** the spearman rank correlation for the same comparison (rho=0.47, *P*=0.00014). A negative slope or a negative rho indicates a codon that is preferred near an exon boundary. For each codon we can then compare these trends at 5' and 3' ends. We note that overall exons tend to have symmetrical trends. The blue line indicates the SMA regression.

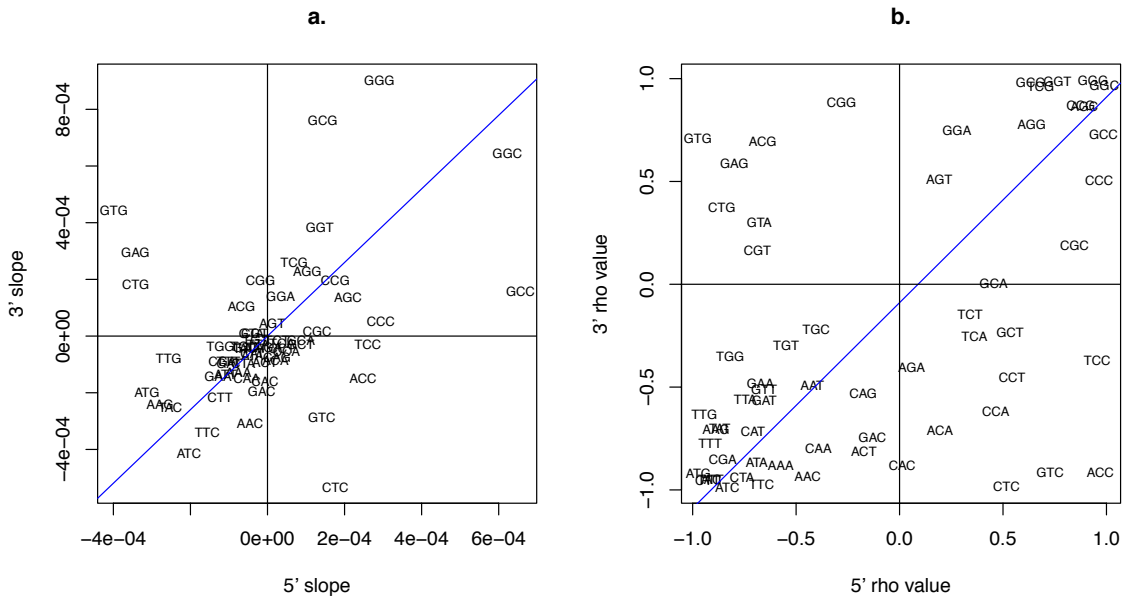


Figure 5. Examination of symmetry of preference/avoidance trends for amino acids

Both 5' and 3' we considered both the slope on the line of relative usage versus distance from the boundary and the spearman rank correlation for the same comparison. For each amino acid we can then compare these trends at 3' and 5' ends, considering either **a.** slope (rho=0.60, *P*=0.003) or **b.** rho (rho=0.68, *P*=0.0005). We note that overall exons tend to have symmetrical trends. The blue line indicates the SMA regression.

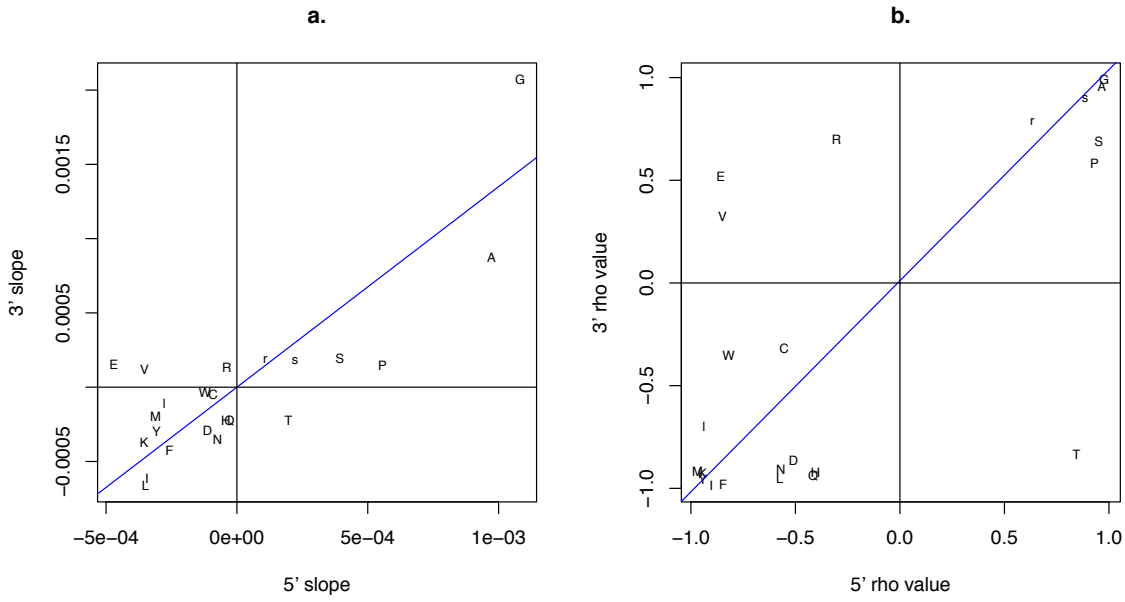


Figure 6. Nucleotide usage at 5' and 3' exon ends at four-fold degenerate sites in *Ectocarpus* and humans. The data here employ only exons longer than 64 codons so that all exons contribute equally at all distances. The plots are **a.** *Ectocarpus* 5' end, **b.** *Ectocarpus* 3' end, **c.** Human 5' end and **d.** Human 3' end.

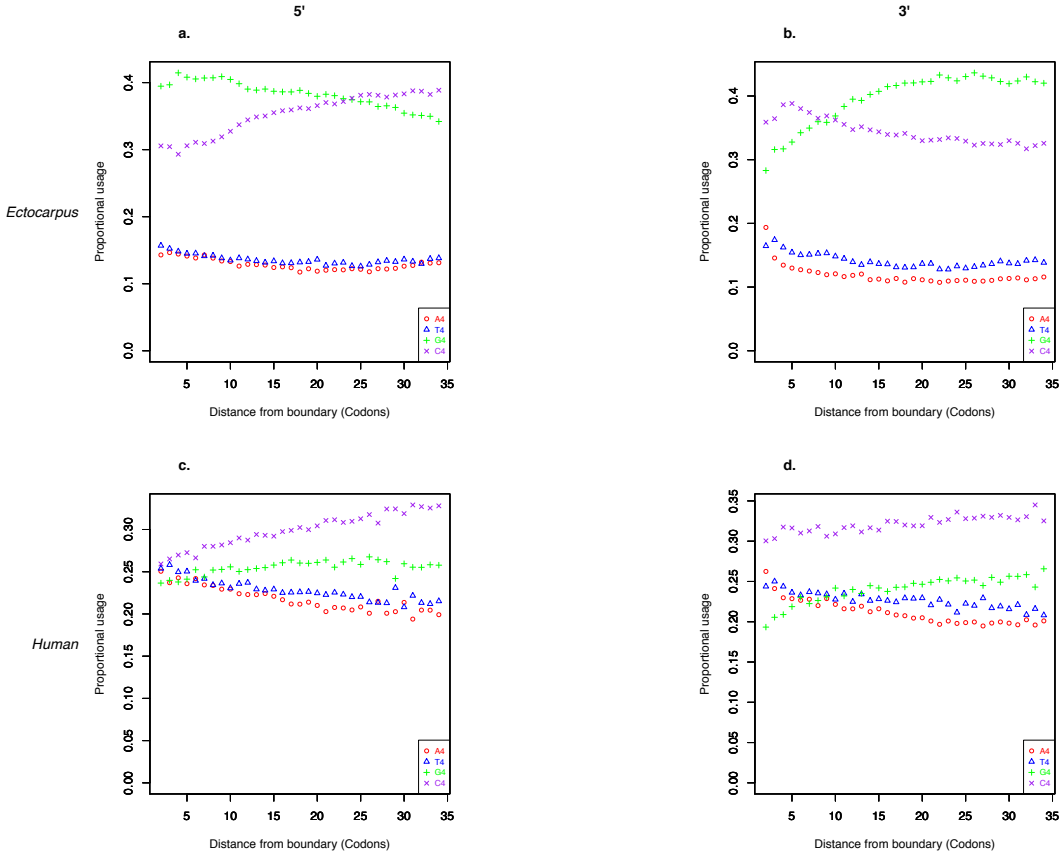


Figure 7. Comparison of preference/avoidance trends at the amino acid level between human and *Ectocarpus*. The amino acid level preference avoidance trends, assayed by rho (the rank correlation of proportional usage of the amino acid to distance from an exon boundary), at **a.** 5' (rho=0.68, *P*=0.0005) and **b.** 3' (rho=0.53, *P*=0.01) ends of exons are shown. The blue line is the SMA regression line.

a. **b.**

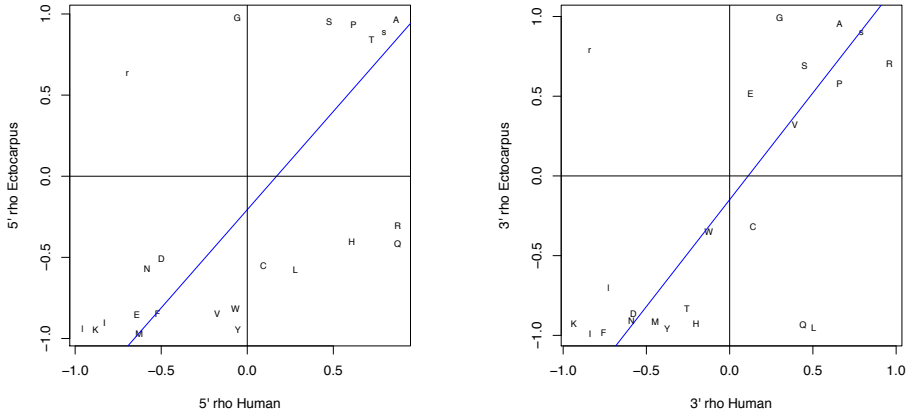


Figure 8. Comparison of preference/avoidance trends at the codon level between human and *Ectocarpus*. The amino acid level preference avoidance trends, assayed by rho (the rank correlation of proportional usage of the amino acid to distance from an exon boundary), at **a.** 5' (rho=0.50, P=5.17 x 10^{-5}) and **b.** 3' (rho=0.58, P=1.7 x 10^{-6}) ends of exons are shown. The blue line is the SMA regression line.

a. **b.**

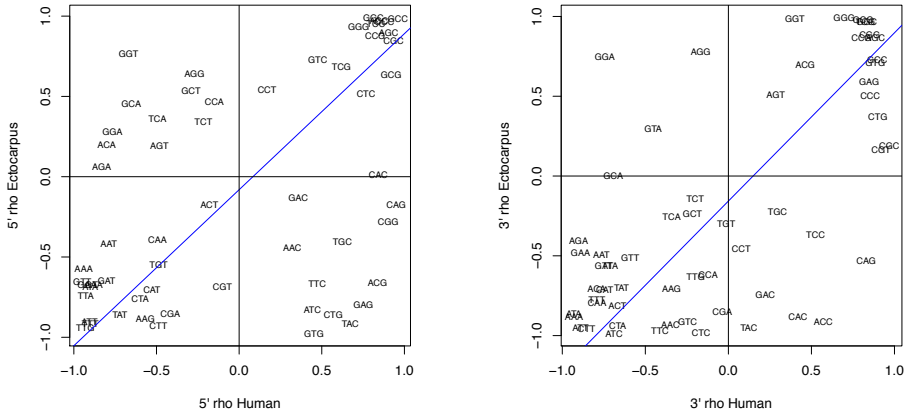


Figure 9. The average number of alternative splicing events detected when genes are grouped according to the number of ESTs aligning to them. Data for humans in red, data for *Ectocarpus* in blue.

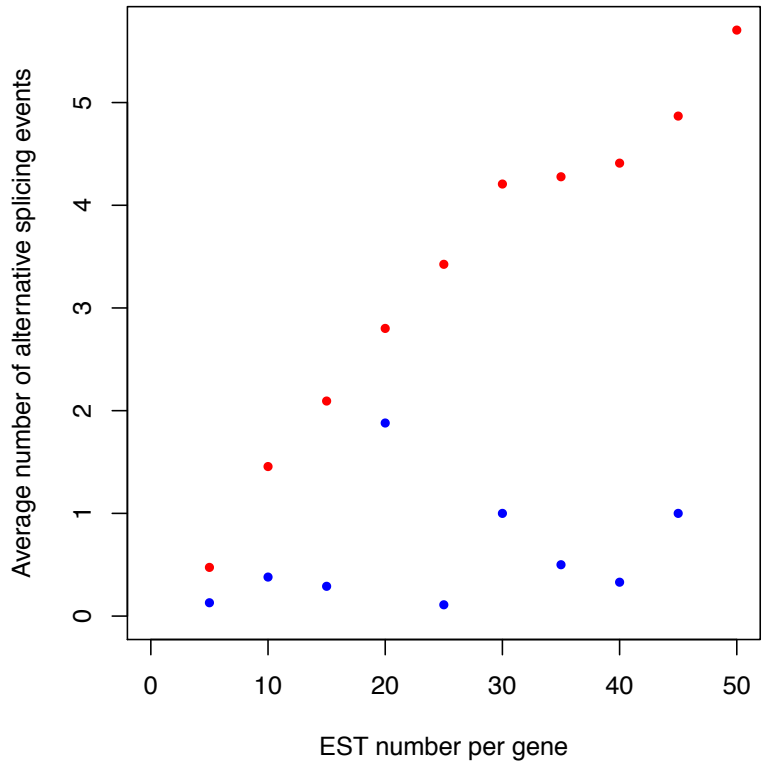
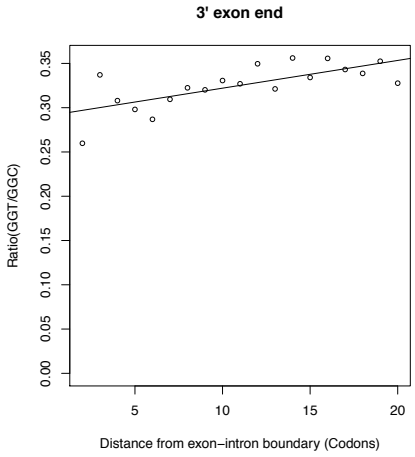
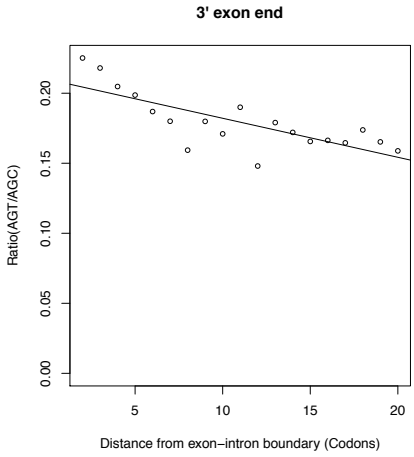


Figure 10. Relative usage of NGT against NGC at synonymous sites at exonic 3' ends a. N=G, **b.** N=A, **c.** N=C, **d.** N=T.

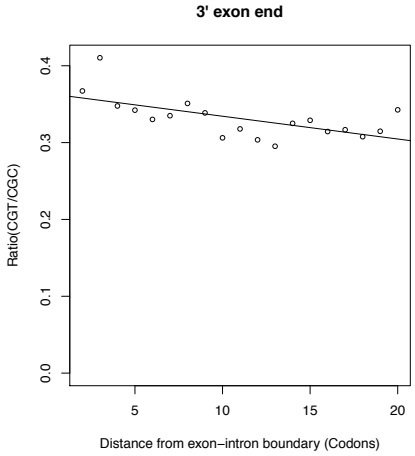
a.



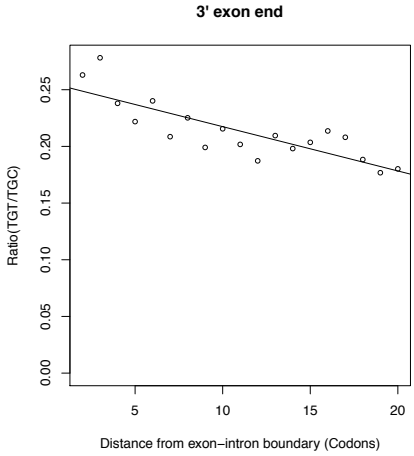
b.



c.



d.



Article – Discoveries

Presence/absence variation in *A. thaliana* is primarily associated with genomic signatures consistent with relaxed selective constraints

Stephen J. Bush[1], Atahualpa Castillo-Morales[1], Jaime M. Tovar-Corona[1], Lu Chen[1,2], Paula X. Kover[1] and Araxi O. Urrutia[1§]

[1]Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK

[2]Present address: Human Genetics, Wellcome Trust Sanger Institute, Genome Campus, Hinxton CB10 1HH, UK

[§]Corresponding author

Email addresses:

SJB: sb444@bath.ac.uk

ACM: acm39@bath.ac.uk

JMTC: jmc21@bath.ac.uk

LC: cl10@sanger.ac.uk

PXK: p.x.kover@bath.ac.uk

AOU: a.urrutia@bath.ac.uk

ABSTRACT

The sequencing of multiple genomes of the same plant species has revealed polymorphic gene and exon loss. Genes associated with disease resistance are over-represented amongst those showing structural variations, suggesting an adaptive role for gene and exon presence/absence variation (PAV). To shed light on the possible functional relevance of polymorphic coding region loss and the mechanisms driving this process, we characterised genes which have lost entire exons or their whole coding regions in 17 fully sequenced *Arabidopsis thaliana* accessions. We found that although a significant enrichment in genes associated with certain functional categories is observed, PAV events are largely restricted to genes with signatures of reduced essentiality: PAV genes tend to be newer additions to the genome, tissue specific and lowly expressed. In addition, PAV genes are located in regions of lower gene density and higher transposable element density. Partial coding region PAV events were associated with only a marginal reduction in gene expression level in the affected accession and occurred in genes with higher levels of alternative splicing in the Col-0 accession. Together these results suggest that although adaptive scenarios cannot be ruled out, PAV events can be explained without invoking them.

INTRODUCTION

Intra-species variation in gene content represents an important source of heterogeneity in the genome of a species and potentially contributes to an organism's adaptability in response to external pressures (Feuk, et al. 2006). Cataloguing significant gains and losses in coding regions within or between species will allow a deeper understanding of the mechanisms underlying the molecular evolution of genomes, and can assist in identifying functional variation in agronomically elite varieties of staple crops (Wang, et al. 2013b). To this end, several studies have examined polymorphic full or partial gene loss in several plant species. For instance, after re-sequencing 50 rice genomes, up to 1327 possible gene loss events (2.4% of the total gene set) were detected relative to the Nipponbare reference accession (Xu, et al. 2012). Significant intra-species variation in gene content has also been reported in maize (Swanson-Wagner, et al. 2010), sorghum (Zheng, et al. 2011) and soybean (McHale, et al. 2012). Previous studies in the model plant *Arabidopsis thaliana*, using re-sequencing microarrays and Illumina sequencing-by-synthesis reads, have also shown significant variations in total nuclear genome sequence among naturally occurring strains (Clark, et al. 2007; Ossowski, et al. 2008). A more recent study using 18 fully sequenced *A. thaliana* genomes found that, relative to the reference accession Col-0, 93.4% of proteins had intra-species variation in their genes, inclusive of large deletions (Gan, et al. 2011) with around 775 genes per accession found to have deletions spanning 50% or more of their coding region sequence (Gan, et al. 2011). A comparison of 80 *Arabidopsis* genomes found that 9% of the total genes in *A. thaliana* showed presence/absence variation (PAV) averaging 444 absent genes per accession (Tan, et al. 2012).

Characterisation of coding region presence/absence variation has shown certain gene categories to be significantly enriched. For instance, 52 of the 154 nucleotide-binding site leucine-rich repeat (NBS-LRR) *R* (resistance) genes were found to be deleted in at least one of fifty rice cultivars (Xu, et al. 2012). Similar over-representation of the *R* genes in *A. thaliana* has also been observed (Bakker, et al. 2006; Shen, et al. 2006), whilst in the soybean, genes enriched in structural variation are more likely to be involved in nucleotide binding and biotic defence (McHale, et al. 2012). Enrichment of particular functional gene categories among genes affected by structural polymorphism suggests these structural polymorphisms may have a functional role, allowing accessions to be better adapted to the environmental conditions they face.

However, this hypothesis has not been explicitly tested. If significant polymorphic deletions are adaptive, we would expect that affected genes should show multiple signatures of being under selection. On the other hand, if structural polymorphisms mostly affect genes evolving under relaxed constraints, then their adaptive significance should be questioned.

Here we characterise genes affected by presence/absence variation (PAV) spanning whole exons in *A. thaliana*, to investigate which genomic features, if any, are associated with these polymorphisms. Our results provide insights into the likely functional impact of structural variation in protein-coding genes.

RESULTS

In order to characterise presence/absence variation (PAV) in *A. thaliana*, we examined previously identified polymorphic deletions in 17 fully sequenced *Arabidopsis* accessions for which transcriptome data was available (Gan, et al. 2011) (see methods). We compiled a set of deletions that spanned entire exons in any of 17 accessions relative to the Col-0 reference genome. A subset of the annotated deletions

was experimentally validated (Gan, et al. 2011). To further rule out the possibility of wrongly identifying deletions due to differences between assemblies, exons were confirmed as missing by searching for homology between the Col-0 exon on all other accessions (see methods).

A total of 794 exons were classified as missing in at least one of 17 accessions, corresponding to 411 genes (approx. 1.5% of the total gene set) including 81 genes where the full coding region was completely absent in at least one accession (supplementary table 1). Exon losses are not uniformly distributed throughout the gene: missing exon sequences are more often found nearer the ends of each gene (supplementary fig. 1).

Overall, approx. 0.3% of the genes in each accession have at least one missing exon, representing between 10-50kb of missing sequence per accession (supplementary table 2). A total of 200 genes had exon loss affecting more than one accession, consistent with a previous study reporting a 'common history' to deletion events in *A. thaliana* (Santuari, et al. 2010).

Because partial deletions spanning whole exons might have distinct functional implications compared to full coding region deletions, the 330 genes with partial coding region loss spanning at least one full exon in at least one accession (exon presence/absence variation, E-PAV) and the 81 genes with full coding region polymorphic deletions affecting at least one accession (full coding DNA sequence presence/absence variation, CDS-PAV) were examined separately.

Genes involved in signal transduction and both nucleotide and protein binding are over represented among PAV genes

In order to characterise PAV genes, we first assessed whether these genes were over-represented in particular gene classes or gene ontology (GO) categories. To do so, we used four classification schemes: 'GO', a condensed set of GO terms (GOslim), the Pfam protein domain database and the family classification scheme of (Gan, et al. 2011) (see methods). Of the 330 E-PAV genes we found most to be poorly characterised with 50% of them having no associated GOslim term. The proportion of poorly characterised genes is greater among CDS-PAV genes, with more than 60% having no associated GOslim term for biological process. When examining genes with associated GOslim terms we found both E-PAV and CDS-PAV genes to be significantly enriched in genes associated with signal transduction and nucleotide binding (fig. 1 and supplementary fig. 2). Furthermore, E-PAV genes also appear significantly enriched in genes associated with the GOslim term 'other binding', which includes proteins that bind to lipids, metal ions and ATP, among other cofactors (fig. 1). Significant over-representation of functional categories among PAV genes is consistent with a previous assessment of large coding region indels in the soybean genome (McHale, et al. 2012) and of whole gene deletions in *A. thaliana* (Tan, et al. 2012). This is also observed when classifying genes using the broader set of 'GO', rather than 'GOslim' terms (supplementary fig. 3).

When classifying genes by family we observe an over-representation of members of the NBS-LRR (nucleotide binding site leucine rich repeat) family – involved in pathogen detection (DeYoung and Innes 2006) – among E-PAV genes (families 'NBS-LRR active TNL', adjusted p-value = 8.57×10^{-35}, and 'NBS-LRR active CNL', adjusted p-value = 4.63×10^{-5}; fig. 1), consistent with previous findings (Shen, et al. 2006). Furthermore, when examining the 3753 Pfam ID gene associations (supplementary figs. 4 and 5) we observe an over-representation of members of the

NB-ARC and LRR domain containing families (note that 'NBS-LRR' refers to a composite of the NBS and LRR domains and that the NBS domain is also known as 'NB-ARC' (McHale, et al. 2006)). No enrichment of any particular gene family was observed among CDS-PAV genes (data not shown).

These significant enrichments in gene functional and domain annotations are in line with previous findings in *Arabidopsis* (Tan, et al. 2012) and other plant species (McHale, et al. 2012; Swanson-Wagner, et al. 2010; Zheng, et al. 2011) and have been proposed to reflect the adaptive role of large polymorphic deletions.

Genes affected by PAV show signatures consistent with relaxed selective constraints

To determine if PAV genes are generally associated with fast evolving proteins potentially under positive selection, we examined the rates of non-synonymous to synonymous changes per gene (dN/dS). Using a randomisation test, E-PAV genes were found to have a significantly higher dN/dS ratio compared to genes with all exons present but only 8 genes have a dN/dS ratio above 1 (fig. 2 and supplementary tables 1 and 3). CDS-PAV genes had a non-significant increase in dN/dS compared to intact genes (those not affected by deletions spanning at least one exon in any accession; fig. 2 and supplementary table 3).

To further examine the selective pressures associated with PAV genes, we examined nucleotide diversity. We considered nucleotide diversity at both replacement sites and silent sites (defined as non-coding sites and the synonymous sites of protein-coding regions) for each gene, according to (Gan, et al. 2011). PAV genes were found to be

associated with higher nucleotide diversity in both silent and replacement sites (supplementary table 3).

While higher dN/dS and nucleotide diversity are suggestive of relaxed selective constraints this pattern is also consistent with a scenario of positive and/or balancing selection. To differentiate between these possible scenarios, Tajima's *D* was calculated for each gene (see methods). A threshold of ±2 was considered as the point at which *D* significantly departs from the null expectation of neutral evolution for any given gene. Of the 330 E-PAV genes with exon presence/absence variation, 24 have $D < -2$ and only 2 have $D > 2$ (AT1G12180, $D = 2.17$, and AT5G35460, $D = 2.05$, both of which are functionally uncharacterized). Among CDS-PAV genes, only 7 have $D < -2$ and none have $D > 2$. Compared to the set of intact genes, there are no significant differences in the proportion of PAV genes either with $D < 2$ (randomisation test $p = 1$ for both E- and CDS-PAV genes) or $D > 2$ (randomisation test $p = 0.93$ and $p = 1$ for E- and CDS-PAV genes, respectively). As demographic characteristics of the *Arabidopsis* population may result in a shift in the average Tajima's *D* among the general pool of genes it is possible that these hard thresholds may not be informative. Indeed, we find that intact genes in *Arabidopsis* have the average Tajima's *D* estimate shifted towards negative values. Thus, PAV genes could fall short of the hard threshold of +2 and still have a higher *D* estimate than the general pool of genes, suggestive of balancing selection. However, E-PAV genes do not show significant differences in Tajima's *D* estimates compared to intact genes and CDS-PAV genes have, in fact, a significantly lower estimate of *D* (fig. 2 and supplementary tables 1 and 3). It is possible that PAV genes may have a higher range of *D* values compared to intact genes, hiding a higher proportion of genes under positive and balancing selection which would not be reflected in overall changes in

the mean. To test this, we compared the distributions of Tajima's *D* estimates in the three sets of genes (intact, E-PAV and CDS-PAV). However, we did not observe any evidence for increased dispersion in *D* among PAV genes (supplementary fig. 6). To further examine this possibility, we examined the proportion of PAV genes below the fifth and above the 95th percentile of the 'intact' distribution (*D* = -2.05 and 1.39, respectively). If a significantly higher proportion of E-PAV or CDS-PAV genes are found compared to the intact set at the positive end of the distribution, we can infer the existence of a detectable subset of PAV genes that may be undergoing balancing selection. However, this is clearly not seen – only 2.33% of E-PAV, and no CDS-PAV genes, exceed the threshold value. At the opposite end of the distribution, we observe no overrepresentation in the proportion of E-PAV genes whose estimates of *D* are lower than the threshold (3.32%) although do observe this for CDS-PAV genes (8.82%). This finding would suggest that a significant proportion of CDS-PAV genes might be undergoing stronger purifying or positive selection relative to intact genes. Together these results suggest that while we cannot rule out the effect of balancing selection acting on a few individual PAV genes a general trend of balancing selection for PAV genes does not readily apply. The excess of negative *D* values among PAV genes coupled with the higher levels of nucleotide diversity and the significant increases in dN/dS ratios are consistent with a scenario of weaker purifying selection but could also be explained by positive selection.

We examined a number of parameters which have been previously associated by some studies with gene essentiality to further explore the functional importance of PAV genes, including a gene's age (Chen, et al. 2012b) and the number of paralogues it has (Hanada, et al. 2009; Makino, et al. 2009), along with weaker associations such as expression level (Cherry 2010) and tissue specificity (Wolf, et al. 2006).

Compared to newer genes, older genes are more likely to be essential (Chen, et al. 2012b). After using the phylogenetic relationships of plant genomes to create a proxy for gene age, we observed that the 330 genes affected by E-PAV are more likely to be newer additions to the genome (fig. 2 and supplementary table 3). It is also possible that E-PAV genes have a greater number of paralogous genes which might compensate for any loss of function. Consistent with this, we find that those genes with missing exons have higher number of paralogues compared to those genes with all exons present (fig. 2 and supplementary table 3). However, the opposite result was observed when analysing CDS-PAV genes – these have an average of 4.2 paralogues compared to genes with no exon losses (fig. 2 and supplementary table 3), suggesting their function is less essential. We then assessed the expression patterns of genes affected by exon presence/absence, since broadly and highly expressed genes are typically associated with higher levels of selection (Yang 2009). Using a randomisation test, we found that genes with exon losses in one or more accessions, when compared to intact genes, had lower expression levels and higher tissue specificity (supplementary table 3). In addition we also observed that exons missing in at least one accession are, on average, shorter than exons present in all accessions (170bp *vs.* 284bp, randomisation test $p = 9.9e^{-5}$; supplementary table 3). However, although exons affected by polymorphic deletions are shorter on average compared to non-deleted exons, E-PAV genes are longer than unaffected genes (2360bp compared to 2142bp, respectively; randomisation test $p = 0.008$; supplementary table 3). By contrast, CDS-PAV genes – where polymorphic deletions encompass the gene's entire coding region – were found to be shorter than unaffected genes (640bp compared to 2142bp, randomisation test $p = 9.9e^{-5}$; supplementary table 3).

Overall, these findings show that although certain functional categories are over-represented among genes with exon loss, more generally significant coding region loss is prevalent amongst novel, lowly expressed and poorly functionally characterised genes. These genes seem to have evolved more recently in the *Arabidopsis* genome and are likely to be under reduced selective constraint.

PAV genes are located in genomic regions that are gene-poor and transposable element-rich

When characterising the genomic context of genes affected by PAV we found that genes with both exon and full coding region loss are separated by longer intergenic distances (fig. 3 and supplementary table 3). Transposable element density around PAV genes was then assessed as gene-poor areas have been associated with a higher transposable element density (Wright, et al. 2003). To do this, we used the reference accession (Col-0) and calculated TE density for each gene in all intergenic sequence in 1-100kb windows centred on each gene's midpoint, by counting the number of bases found within TE annotations (see methods). E-PAV genes were found to have an approximately two-fold increase in the amount of bases annotated as a TE compared to genes which are intact in all accessions (e.g. TE sequence accounts for approx. 30% of the non-genic sequence within a 10kb window surrounding an E-PAV gene; fig. 3 and supplementary table 4). Significant enrichment of specific transposable element superfamilies was also observed, notably DNA transposons and LTR retrotransposons (supplementary table 4).

In addition, we found that genes with missing exons have, on average, a shorter distance from the gene boundary to the nearest TE than those genes with all exons present (2.5kb compared to 5.7kb; randomisation test, $p = 9.9e^{-5}$; supplementary table

3. If calculating the minimum distance to the nearest TE, classified by superfamily, E-PAV genes are significantly closer to every TE type: rolling circle TEs, DNA transposons, LTR retrotransposons, LINEs and SINEs (supplementary table 3). Similar findings were obtained when analysing TE content in the surrounding regions of CDS-PAV genes (supplementary table 3).

Certain TE sequence motifs have been associated with recombination hotspots which could drive exon loss through promoting ectopic recombination events (Horton, et al. 2012; Oliver and Greene 2009). To explore whether genes affected by PAV have a local enrichment for such hotspot motifs, we examined the density of these motifs both in and around genes (see methods). However, we observed no significant differences in hotspot motif occupancy in the non-genic regions of windows surrounding E-PAV genes compared to intact genes (in window sizes of 1kb to 100kb centred on the gene's midpoint; supplementary table 4). Nevertheless, a significant enrichment in hotspot motif occupancy was observed in the genic sequence of all windows centred on E-PAV genes compared to those centred on 'intact' genes (fig. 3 and supplementary table 4). When comparing CDS-PAV genes to the intact set, we observed no consistent pattern of higher hotspot motif density within genic regions and only a marginally higher proportion of hotspot motifs in the non-genic regions that surround them, in windows up to 3kb in size ($p < 0.01$; supplementary table 5). Taken together, these results show that PAV genes are located in gene-poor and TE-rich regions of the genome further supporting the hypothesis that PAV is associated with relaxed selective constraints. Enrichments of sequence motifs previously associated with recombination hotspots in or around PAV genes suggest that at least some exon deletion events may have resulted from recombination events involving these recombination hotspot motifs.

Exon loss is associated with a marginal reduction in expression level

The above results suggest that exon presence/absence variation is associated with reduced selective constraints. To assess whether exon loss is likely to have resulted in reduced functionality for the genes affected, we compared expression levels for genes with and without missing exons across accessions. If exon loss causes or follows from diminished functionality by previous mutations we would expect expression to be significantly reduced in those accessions affected by E-PAV. Using RNAseq transcription profiles for each *Arabidopsis* accession (Gan, et al. 2011), we compared the expression patterns of individual genes in accessions affected by exon deletions with those accessions where the gene remained intact. To do this, we transformed expression data per accession to Z-scores (Cheadle, et al. 2003). We then looked only at those genes where exon loss had occurred in a single accession (210 genes). For each gene, we took (a) the expression level of that gene in the affected accession, and (b) the mean expression level of that gene across the 17 unaffected accessions (the other 16 under study plus the reference genome, Col-0). We found that half of the genes examined had an expression level below this mean and 37% an expression level equal to it. However, on average, expression levels in the affected accession departed little from mean expression in unaffected accessions (0.15 standard deviations). In 27 genes (13% of cases), expression level in a gene affected by an exon deletion was higher than the mean expression across unaffected accessions with 14 cases showing a statistically significant difference (fig. 4 and supplementary table 6). These 27 genes are generally poorly characterised with 12 having no functional category annotations. Most genes affected by exon deletions had low expression levels to begin with, although some exceptions are notable, such as rotamase CYP4 (AT3G62030;

involved in a variety of cellular functions related to metabolism and response to several types of stress), which has an average expression level in the unaffected accessions of 400rpkm, among the top 1% of genes with detectable expression in Col-0.

It is possible that the moderate effect of exon loss on gene expression levels is explained by an over-representation of alternatively spliced exons amongst the set of missing exons. This would allow for the production of viable protein products in their absence. In order to test this, we quantified alternative splicing in 15,540 *Arabidopsis* genes including 103 of the 330 E-PAV associated genes using a 'comparable alternative splicing index' (see methods) which corrects for the distorting effect of variation in transcript coverage among genes (reviewed in (Chen, et al. 2012a)). E-PAV genes were found to have a significantly higher number of alternative splicing events compared to intact genes (3.35 and 1.13 respectively; randomisation test $p = 0.046$).

Overall, these findings suggest that exon losses have only a marginal effect on the expression profile of genes in the accessions affected. The higher levels of alternative splicing among genes affected by exon loss raises the possibility that a significant proportion of lost exons are normally alternatively spliced, reducing selection pressure on these exons since a functional protein product would be produced in their absence anyway.

DISCUSSION

Intra-species structural variations in genes have been proposed to play an important role in the adaptation of particular populations to variation in environmental conditions (Feuk, et al. 2006). Here we have characterised presence/absence coding sequence variation (PAV) in 17 fully sequenced *A. thaliana* genomes, relative to the

reference accession Col-0, affecting 411 genes including 81 instances of whole coding region deletions. We found a significant enrichment of genes associated with the GO terms for protein and nucleotide binding as well as signal transduction. Both gene family and Pfam annotation enrichment analysis revealed significant enrichments of gene members from the disease resistance associated NBS-LRR gene families. Significant deviations from random expectations have been observed in previous studies of PAV genes in plants, with similar over-representation of resistance-associated gene families among PAV genes. For instance, in sorghum (Zheng, et al. 2011), PAV genes are enriched in nine Pfam categories, including the NB-ARC domain-containing family. In soybean (McHale, et al. 2012), PAV-affected genes have also been found to be enriched for members of the NB-ARC family, and within the GO category of 'defence response.' CDS-PAV genes have also been shown to deviate from random expectations in *Arabidopsis* (Tan, et al. 2012), with the greatest significant enrichment in PAV genes also reported for those with NB-ARC domains. These functional and/or gene family enrichments can be suggestive of an adaptive role for PAV events by aiding specific ecotypes in adapting to their local environment. Our results – showing that genes associated with, e.g., resistance are more likely to be affected by PAV – are, at first glance, consistent with this hypothesis. In addition, we were able to confirm a previous report of CDS-PAV for three members of the *R* gene family – the single-exon gene AT5G05400, and the multi-exon genes AT5G18350 and AT5G49140 (Shen, et al. 2006) – a family known to have signatures of positive selection in *A. thaliana* (Mondragon-Palomino, et al. 2002). However, comprehensive analysis for evidence of selection does not support this as a general interpretation.

dN/dS ratios are one of the most widely used estimates of selective pressure acting on protein coding genes with dN/dS >> 1 indicative but not a definitive signature of positive selection (Hurst 2002). Although there are, on average, a higher number of substitutions in E-PAV genes compared to intact genes, this is not a clear signature of adaptation, and can suggest comparatively relaxed negative, rather than stronger positive, selection.

We further found that PAV genes have significantly higher nucleotide diversity both at silent and replacement sites. Both observations are suggestive of weaker purifying selection; however, they can also be expected if PAV genes were under higher balancing selection. Indeed, there is evidence to suggest that the diversity of resistance-associated genes is maintained by balancing selection (Van der Hoorn, et al. 2002), which are over-represented among PAV genes. Balancing selection has been proposed to stably maintain both the intact gene and the absent allele (Tan, et al. 2012).

So, is balancing selection the most parsimonious explanation for why PAV genes are associated with higher nucleotide diversity? A classic scenario of trans-species polymorphism, associated with balancing selection, cannot be assessed given the limited sequence variation data available for *A. lyrata*, *A. thaliana*'s closest sequenced relative. It is possible that the 'gene/exon present' and the 'gene /exon absent' alleles are under selection to be maintained in different *A. thaliana* populations, allowing them to better adapt to their local environment. This would be consistent with the increase in nucleotide diversity but this scenario cannot be distinguished from alternative neutral models. Conditional neutrality at PAV loci, where the functional gene has ceased to be adaptive in some but not all environments, cannot be ruled out (e.g. in the case of resistance genes where the corresponding pathogen is absent (Gos

and Wright 2008)). In this case, the absent allele would have no selective advantage at any point but rather result from relaxed constraints associated with PAV genes in some *Arabidopsis* populations. Moreover, a model of generalised relaxed constraints affecting the PAV loci would also lead to increased nucleotide diversity and slight increases in dN/dS.

Tajima's *D*, a comparison of two estimators of θ (the population mutation rate $4Ne\mu$) – the number of segregating sites and the average number of pairwise differences between sequences (Tajima 1989) – offers a more reliable estimate of selective pressures acting on a gene as it incorporates information about the distribution of segregating alleles in a species. This allows more accurate estimations of the degree and direction of departure of sequence evolution from a neutral expectation (although non-selectionist interpretations of *D* are also possible, such as recent population expansion or bottlenecking for negative and positive *D*, respectively) (Tajima 1989). Tajima's D values do not provide evidence for either E-PAV or CDS-PAV genes to be under balancing selection. Taking dN/dS, nucleotide diversity and *D* estimates together, most PAV genes appear to be evolving under relaxed constraints.

A signature of relaxed selection associated with PAV genes is combined with a variety of features which have been associated with lower gene essentiality. We found that PAV genes have lower expression levels and higher tissue specificity; both of these features have been associated with higher rates of substitutions and reduced gene essentiality (Cherry 2010; Wolf, et al. 2006). Older genes have been considered more essential (Chen, et al. 2012b) and have been associated (in humans, flies and *Aspergillus*) with a higher expression level and stronger purifying selection (Wolf, et al. 2009). We found that PAV genes are, on average, newer additions to the genome and that most exons affected by PAV do not have an orthologous exon in *A. lyrata*

(663/794). We note that both E-PAV and CDS-PAV genes are enriched in reverse transcriptase domains (supplementary figs. 4 and 5) and E-PAV genes for transposase domains (supplementary fig. 4), suggesting exonization of transposable elements as the origin of some PAV-affected exons.

In addition, the fact that gene expression is only marginally reduced in accessions affected by exon deletion events suggests that the lost exons may only have had a limited impact on gene functionality. This is possibly explained in some cases by alternative splicing, which has already been associated with an increased frequency of exon loss in humans, mice and rats – alternatively spliced forms are less likely to be conserved between species than constitutive exons (Modrek and Lee 2003). In *A. thaliana*, we found that genes with E-PAV are under weaker purifying selection and have a greater number of alternative splice events compared to intact genes. This observation suggests that alternatively spliced exons are likely to be under reduced selective constraints compared to constitutive exons and thus whole exon deletions would have less of a detrimental effect than the loss of a constitutive exon. To the best of our knowledge this is the first time that exon loss events have been associated with elevated alternative splicing levels within a species rather than between species.

The genomic context of genes has also been linked to both patterns of sequence evolution and features associated with gene essentiality. A recent study in *A. thaliana* has correlated the presence of TEs adjacent to genes with sequence variation within that gene (Wang, et al. 2013a) suggesting TEs tend to accumulate near genes under lower selective pressures located in regions with less efficient purging of TE sequence. Indeed, for our set of E-PAV genes, we find a higher density of TEs in the vicinity. In addition, we also find that genes undergoing PAV have an increased proportion of motifs associated with recombination hotspots within their sequence.

Both findings are consistent with PAV events being associated with genes located in genomic regions evolving under reduced selective constraints. Moreover, higher TE content and hotspot motifs are consistent with the suggestion that unequal recombination between homologues may be a major mechanism for generating P/A polymorphisms (Tan, et al. 2012). However, it should be noted that no recombinogenic motif is both necessary and sufficient for a recombination event to occur (Johnston and Cutler 2012) and as such, their connection, if any, to PAV remains speculative.

All of these features considered together suggest that although some individual deletions might have an adaptive value, overall coding region loss disproportionally affects genes under reduced selective pressures. So how are these results reconciled with the enrichment of certain gene families and GO functional terms? The enrichment of specific functional categories and gene families among PAV genes (fig. 1), leads to the implication of adaptive pressures favouring PAV on genes related to specific biological processes (Tan, et al. 2012). However, as we have shown, PAV genes are associated with a variety of features suggestive of lower selective constraints. We argue that the enrichment of certain GO categories and/or gene families among genes associated with a particular genomic feature does not, by itself, allow us to draw conclusions about any adaptive processes these genes may be undergoing. Consistent with this, we find that intact genes associated with the gene categories in which PAV genes are enriched, also show the same signatures of reduced selection (supplementary table 7). This is notable for those sets of genes involved in, e.g., signal transduction, nucleic acid binding and the NBS-LRR family – categories enriched among PAV genes (fig. 1). For instance, if we compare the set of E-PAV genes to the set of genes with all exons present, and the set of NBS-LRR

genes to the set of genes belonging to other families, we find that both E-PAV and NBS-LRR genes are comparatively newer additions to the genome, have a higher dN/dS ratio, a higher number of alternative splicing events, a higher number of paralogues, a higher proportion of SNPs and are found closer to TEs (supplementary table 7). We note that the proportion of polymorphic sites is higher not only in PAV genes but in genes of that functional category. To demonstrate that PAV genes do not bias the comparison of, e.g., the set of NBS-LRR genes to the set of genes belonging to other families, we repeat the analysis restricted to intact genes only and observe the same result (supplementary table 7).

The fact that we observed fewer PAV genes than a previous study examining 80 fully sequenced *Arabidopsis* genomes (n = 2741 (Tan, et al. 2012)) is likely due to differences in methodology. Firstly, our analysis uses 17 genomes assembled using a combination of read-to-reference genome (Col-0) alignment and *de novo* approaches, and – importantly – for which transcriptome data was available (Gan, et al. 2011), rather than the 80 accessions reported by (Cao, et al. 2011). Secondly, we use a more conservative methodology for defining significant deletions whilst (Tan, et al. 2012) define PAV genes using what is referred to as the 'broad definition': "one being found at a particular locus only in some genomes compared to the others." This allows a gene to be called as a PAV gene even if a copy exists at a different locus. To minimise the inclusion of rearrangement events as deletions, (Tan, et al. 2012) examined their predicted PAV genes using blastn against a reference accession, excluding from the 'absent' category any gene with a counterpart that matches >50% of its length. Our definition of PAV is more restrictive as we only deemed an exon or gene to be deleted if genome alignments showed that the deletion spanned at least a whole exon or whole gene with not a single identifiable base remaining. Finally, the

(Tan, et al. 2012) study used genomes assembled according to the TAIR8 annotated positions whereas our data is assembled according to TAIR10. There is a small risk, therefore, of having incorporated now-obsolete gene models into their findings. Regardless of the methodological differences and the resulting variation in sample size it is worth noting our results are not in contradiction to those of previous studies examining PAV both in *Arabidopsis* and other species as we find similar deviations from random expectations in the functional annotations of genes. Our analysis of sequence evolution and other genic features of PAV genes do not rule out the possibilities of conditional neutrality at PAV loci or that balancing selection may be acting on PAV genes, allowing adaptation to the environmental conditions of specific ecotypes. Instead, the findings presented show that PAV events can be explained by a non-adaptive interpretation where genes under reduced constraints are more susceptible to the spread of allele variants containing significant deletions.

In summary, our results suggest that although significant enrichment in functional categories among PAV genes was observed, most exon loss events are observed in newer, poorly functionally characterised genes associated with signatures linked to less essential genes evolving under lower purifying or balancing selection. This may reduce the potential functional relevance of structural variations within these genes. We conclude that while an adaptive model for PAV cannot be ruled out, the observed functional enrichments among PAV genes and increased nucleotide diversity can also be interpreted without invoking selection.

MATERIALS AND METHODS

Genome sequence and annotations. Exon coordinates for *A. thaliana* strain Col-0 were obtained from The Arabidopsis Information Resource (TAIR) (file 'TAIR10_GFF3_genes.gff', dated 20th March 2012). The genomes of 17 *A. thaliana*

accessions (Bur-0, Can-0, Ct-1, Edi-0, Hi-0, Kn-0, Ler-0, Mt-0, No-0, Oy-0, Rsch-4, Sf-2, Tsu-0, Wil-2, Ws-0, Wu-0 and Zu-0) were obtained from (Gan, et al. 2011). We did not use data from Po-0 because it has an unusually high frequency of heterozygosity and high similarity to Oy-0 (Gan, et al. 2011). Each genome has been fully sequenced and assembled, using a combination of *de novo* assembly and read mapping to the reference accession, Col-0.

Detecting missing exons relative to Col-0. For this analysis we selected a set of deletions spanning at least one full exon in at least one accession relative to the Col-0 reference genome from a wider set of deletion events described by Gan et al. (Gan, et al. 2011). Exons absent in the Col-0 reference genome but present in other accessions are not included in any analysis. Confirmation of these deletions is described by the original authors who analysed deletion breakpoints (Gan, et al. 2011). In this dataset, deletion breakpoints were estimated to within ~30bp, with left and right consensus sequences established by growing inwards from these estimates using the read mapping information. If there was a deletion, these two ends would overlap. Gan et al. (Gan, et al. 2011) confirmed this with alignments of the left and right consensus sequences, thus excluding errors of sequencing or misassembly. We further confirmed the presence or absence of each individual exon in each of 17 accessions relative to the Col-0 genome annotation using blastn with default parameters (Altschul, et al. 1990). Sequence alignments were obtained using the best hit homologue and the Smith-Waterman algorithm (fasta35 with parameters –a –A) (Pearson 1999). We confirmed an exon as missing if both (a) an alignment could not be made, and (b) if none of the nucleotide positions in the Col-0 exon mapped to any nucleotide in the accession.

Functional category enrichment analysis. Four gene classification schemes were obtained. 'GOslim' terms were obtained from The Arabidopsis Information Resource (TAIR) (file 'ATH_GO_GOSLIM.txt', dated 9th July 2013), excluding terms unsupported by experimental or computational analysis, i.e. evidence codes ND, NR and NAS. 'GO' term annotations were obtained from Ensembl BioMart (17th July 2013) (Smedley, et al. 2009). 'Pfam' terms were obtained from Pfam v27.0 (17th July 2013) (Punta, et al. 2012). In addition, 7119 genes were classified into 49 distinct families as in (Gan, et al. 2011). Statistical significance of the enrichment of both GOslim, GO terms, of Pfam class and family membership among both E-PAV and CDS-PAV-affected genes was assessed using a Monte Carlo random sampling (1000 randomisations), with the p-value of the enrichment of each category obtained using a Z-test. The significance of individual categories was corrected for multiple testing by the Benjamini-Hochberg procedure.

Sequence evolution analysis. To approximate selective constraint on a gene, we calculated dN/dS. For each gene, we obtained a local alignment of the Col-0 CDS against its *A. lyrata* orthologue, using the Smith-Waterman algorithm (fasta35 with parameters –a –A) (Pearson 1999). dN/dS was calculated using the Yang and Nielson model, as implemented in the yn00 package of PAML (Yang 2009). Using substitution estimates, as above, and SNP data from (Gan, et al. 2011), we also estimated Tajima's *D* (Tajima 1989) per gene. Nucleotide diversity is calculated according to (Gan, et al. 2011).

Paralogue number and gene age annotations. Orthologue and paralogue data were obtained from BioMart (Vilella, et al. 2009). A proxy for gene age was established using taxonomic classifications, based on the phylostratigraphic method of (Domazet-Lošo, et al. 2007). If a candidate orthologue was identified for each *A. thaliana* gene

in any of 15 plant and algal species at a minimum identity of 30%, the gene was considered to be as old as the 'broadest' taxonomic category held in common (see supplementary table 8). This allowed us to make use of orthologue data despite divergence times relative to *A. thaliana* being known for only its closest relatives – at approximately 5 million years for *A. lyrata* (Kuittinen, et al. 2004), and 20 million years for *Brassica rapa* (Yang, et al. 1999).

Gene expression. Expression specificity was calculated as a tissue specificity index (*tau*) (Yanai, et al. 2005), using the parallel signature sequencing (MPSS) database (Brenner, et al. 2000; Meyers, et al. 2004; Nakano, et al. 2006). Expression levels were calculated using RNAseq transcript abundance data, as absolute read values corrected by sequence length in each accession (known as rpkm values: per gene, the number of reads per kilobase per million mapped reads) (Gan, et al. 2011).

Transposable element and hotspot motif density. Transposable element (TE) coordinates for *A. thaliana* strain Col-0 were obtained from The Arabidopsis Information Resource (TAIR) (file 'TAIR10_Transposable_Elements.txt', dated 20th March 2012). For our analyses, we identified every instance of all 25 hotspot-associated motifs (of 5 to 9bp) described by (Horton, et al. 2012) in the Col-0 reference genome. TE and hotspot motif density for each gene was calculated as the proportion of base pairs occupied by a TE or a hotspot motif within windows of size 1-100kb centred on the nucleotide at the gene's midpoint. Windows consist of both coding and non-coding sequence within a region of length (window size)/2 up- and downstream of the midpoint base. Both TE and hotspot motif density were calculated as the number of TE or motif bases, respectively, relative to the number of intergenic or genic bases contained within the window, rather than the total number of bases in the window.

Alternative splicing events. Alternative splicing events were identified using the methods described in (Chen, et al. 2011). In brief, the number of alternative splicing events per gene were identified by aligning EST data obtained from dbEST (Boguski, et al. 1993) to the genome sequence (ftp://ftp.ncbi.nih.gov/repository/dbEST, downloaded 1st May 2011). Those ESTs aligning to regions with no annotated gene were excluded from the analysis. EST alignments were then used to create an exon template. Alternative splicing events per gene were identified by comparing alignment coordinates for each individual EST to exon annotations. As a low EST coverage can increase the number of falsely positive claims that an exon is constitutive, rather than spliced, we excluded genes with 10 or fewer ESTs. ESTs were assigned to genes using gene annotation coordinates. A comparable alternative splicing index that avoids transcript coverage biases was obtained using the transcript normalisation method described in (Kim, et al. 2007). Briefly, for each gene one hundred random samples of 10 ESTs were selected. Finally, the number of alternative splicing events were calculated for each random sample (as detailed above), with an overall average calculated per gene.

Randomisation test. A randomisation test was used to obtain numerical p values to assess the statistical significance of any variation in the characteristics of PAV-affected genes compared to 'intact' genes. In brief, we contrasted genomic feature parameters in E-PAV (n=330) or CDS-PAV genes (n=81) to the distribution of means of the same genomic feature in s=10,000 randomly generated subsets of an equal number of genes drawn from the complete gene set. The numerical p value was calculated as follows: let q be the number of times the mean value of the PAV set exceeded the mean value of the randomly generated subset. Letting $r = s\text{-}q$, then the p-value of this test is $r+1/s+1$.

ACKNOWLEDGEMENTS AND FUNDING INFORMATION

The authors wish to acknowledge the valuable comments made by two anonymous reviewers. This work was supported by a University of Bath fee studentship to SJB, CONACyT scholarships to ACM and JMTC, a UK-China scholarship for excellence and University of Bath research studentship to LC, a BBSRC grant (grant number BB/F022697/1) to PXK and a Royal Society Dorothy Hodgkin Research Fellowship (DH071902), Royal Society research grant (grant number RG0870644) and a Royal Society research grant for fellows (grant number RG080272) to AUO.

REFERENCES

The Arabidopsis Information Resource - *A. thaliana* Genome Version 10 (TAIR10) Blast Datasets [Internet]. Available from: ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/TAIR10_blastsets/

The Arabidopsis Information Resource - *A. thaliana* Genome Version 10 (TAIR10) Gene Ontologies [Internet]. Available from: ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ 1990. Basic local alignment search tool. J Mol Biol 215: 403-410.

Bakker EG, Toomajian C, Kreitman M, Bergelson J 2006. A genome-wide survey of R gene polymorphisms in Arabidopsis. Plant Cell 18: 1803-1818.

Boguski MS, Lowe TMJ, Tolstoshev CM 1993. dbEST - database for expressed sequence tags. Nat Genet 4: 332-333.

Brenner S, Johnson M, Bridgham J, et al. (24 co-authors). 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18: 630-634.

Cao J, Schneeberger K, Ossowski S, et al. (17 co-authors). 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet 43: 956-963.

Cheadle C, Vawter MP, Freed WJ, Becker KG 2003. Analysis of microarray data using Z score transformation. J Mol Diagn 5: 73-81.

Chen L, Tovar-Corona JM, Urrutia AO 2012a. Alternative Splicing: A Potential Source of Functional Innovation in the Eukaryotic Genome. International Journal of Evolutionary Biology 2012: 10.

Chen L, Tovar-Corona JM, Urrutia AO 2011. Increased levels of noisy splicing in cancers, but not for oncogene-derived transcripts. Human Molecular Genetics 20: 4422-4429.

Chen W-H, Trachana K, Lercher MJ, Bork P 2012b. Younger Genes Are Less Likely to Be Essential than Older Genes, and Duplicates Are Less Likely to Be Essential than Singletons of the Same Age. Molecular Biology and Evolution 29: 1703-1706.

Cherry JL 2010. Expression level, evolutionary rate, and the cost of expression. Genome Biol Evol 2: 757-769.

Clark RM, Schweikert G, Toomajian C, et al. (18 co-authors). 2007. Common Sequence Polymorphisms Shaping Genetic Diversity in Arabidopsis thaliana. Science 317: 338-342.

DeYoung BJ, Innes RW 2006. Plant NBS-LRR proteins in pathogen sensing and host defense. Nat Immunol 7: 1243-1249.

Domazet-Lošo T, Brajković J, Tautz D 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends in Genetics 23: 533-539.

Feuk L, Carson AR, Scherer SW 2006. Structural variation in the human genome. Nat Rev Genet 7: 85-97.

Gan X, Stegle O, Behr J, et al. (23 co-authors). 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature 477: 419-423.

Gos G, Wright SI 2008. Conditional neutrality at two adjacent NBS-LRR disease resistance loci in natural populations of Arabidopsis lyrata. Mol Ecol 17: 4953-4962.

Hanada K, Kuromori T, Myouga F, Toyoda T, Li WH, Shinozaki K 2009. Evolutionary persistence of functional compensation by duplicate genes in Arabidopsis. Genome Biol Evol 1: 409-414.

Horton MW, Hancock AM, Huang YS, et al. (13 co-authors). 2012. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. Nat Genet 44: 212-216.
Hurst LD 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet 18: 486.
Johnston Henry R, Cutler David J 2012. Population Demographic History Can Cause the Appearance of Recombination Hotspots. The American Journal of Human Genetics 90: 774-783.
Kim E, Magen A, Ast G 2007. Different levels of alternative splicing among eukaryotes. Nucleic Acids Research 35: 125-131.
Kuittinen H, de Haan AA, Vogl C, Oikarinen S, Leppala J, Koch M, Mitchell-Olds T, Langley CH, Savolainen O 2004. Comparing the linkage maps of the close relatives Arabidopsis lyrata and A. thaliana. Genetics 168: 1575-1584.
Makino T, Hokamp K, McLysaght A 2009. The complex relationship of gene duplication and essentiality. Trends Genet 25: 152-155.
McHale L, Tan X, Koehl P, Michelmore R 2006. Plant NBS-LRR proteins: adaptable guards. Genome Biology 7: 212.
McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL, Gerhardt DJ, Jeddeloh JA, Stupar RM 2012. Structural variants in the soybean genome localize to clusters of biotic stress-response genes. Plant Physiol 159: 1295-1308.
Meyers BC, Tej SS, Vu TH, Haudenschild CD, Agrawal V, Edberg SB, Ghazal H, Decola S 2004. The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. Genome Res 14: 1641-1653.
Modrek B, Lee CJ 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet 34: 177-180.
Mondragon-Palomino M, Meyers BC, Michelmore RW, Gaut BS 2002. Patterns of positive selection in the complete NBS-LRR gene family of Arabidopsis thaliana. Genome Res 12: 1305-1315.
Nakano M, Nobuta K, Vemaraju K, Tej SS, Skogen JW, Meyers BC 2006. Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Research 34: D731-D735.
Oliver KR, Greene WK 2009. Transposable elements: powerful facilitators of evolution. Bioessays 31: 703-714.
Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D 2008. Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Res 18: 2024-2033.
Pearson WR. 1999. Flexible Sequence Similarity Searching with the FASTA3 Program Package. In. p. 185-219.
Punta M, Coggill PC, Eberhardt RY, et al. (16 co-authors). 2012. The Pfam protein families database. Nucleic Acids Research 40: D290-D301.
Santuari L, Pradervand S, Amiguet-Vercher A-M, Thomas J, Dorcey E, Harshman K, Xenarios I, Juenger T, Hardtke C 2010. Substantial deletion overlap among divergent Arabidopsis genomes revealed by intersection of short reads and tiling arrays. Genome Biology 11: R4.
Shen J, Araki H, Chen L, Chen JQ, Tian D 2006. Unique evolutionary mechanism in R-genes under the presence/absence polymorphism in Arabidopsis thaliana. Genetics 172: 1243-1250.
Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A 2009. BioMart - biological queries made easy. BMC Genomics 10: 22.

Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, Springer NM 2010. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res 20: 1689-1699.
Tajima F 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.
Tan S, Zhong Y, Hou H, Yang S, Tian D 2012. Variation of presence/absence genes among Arabidopsis populations. BMC Evolutionary Biology 12: 86.
Van der Hoorn RAL, De Wit PJGM, Joosten MHAJ 2002. Balancing selection favors guarding resistance proteins. Trends in Plant Science 7: 67-71.
Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19: 327-335.
Wang X, Weigel D, Smith LM 2013a. Transposon Variants and Their Effects on Gene Expression in *Arabidopsis*. PLoS Genet 9: e1003255.
Wang Y, You FM, Lazo GR, Luo M-C, Thilmony R, Gordon S, Kianian SF, Gu YQ 2013b. PIECE: a database for plant gene structure comparison and evolution. Nucleic Acids Research 41: D1159-D1166.
Wolf YI, Carmel L, Koonin EV 2006. Unifying measures of gene function and evolution. Proceedings of the Royal Society B: Biological Sciences 273: 1507-1515.
Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proceedings of the National Academy of Sciences 106: 7273-7280.
Wright SI, Agrawal N, Bureau TE 2003. Effects of recombination rate and gene density on transposable element distributions in Arabidopsis thaliana. Genome Res 13: 1897-1903.
Xu X, Liu X, Ge S, et al. (25 co-authors). 2012. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotech 30: 105-111.
Yanai I, Benjamin H, Shmoish M, et al. (12 co-authors). 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21: 650-659.
Yang H 2009. In plants, expression breadth and expression level distinctly and non-linearly correlate with gene structure. Biology Direct 4: 45.
Yang YW, Lai KN, Tai PY, Li WH 1999. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. J Mol Evol 48: 597-604.
Zheng L-Y, Guo X-S, He B, et al. (11 co-authors). 2011. Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). Genome Biology 12: R114.

FIGURES

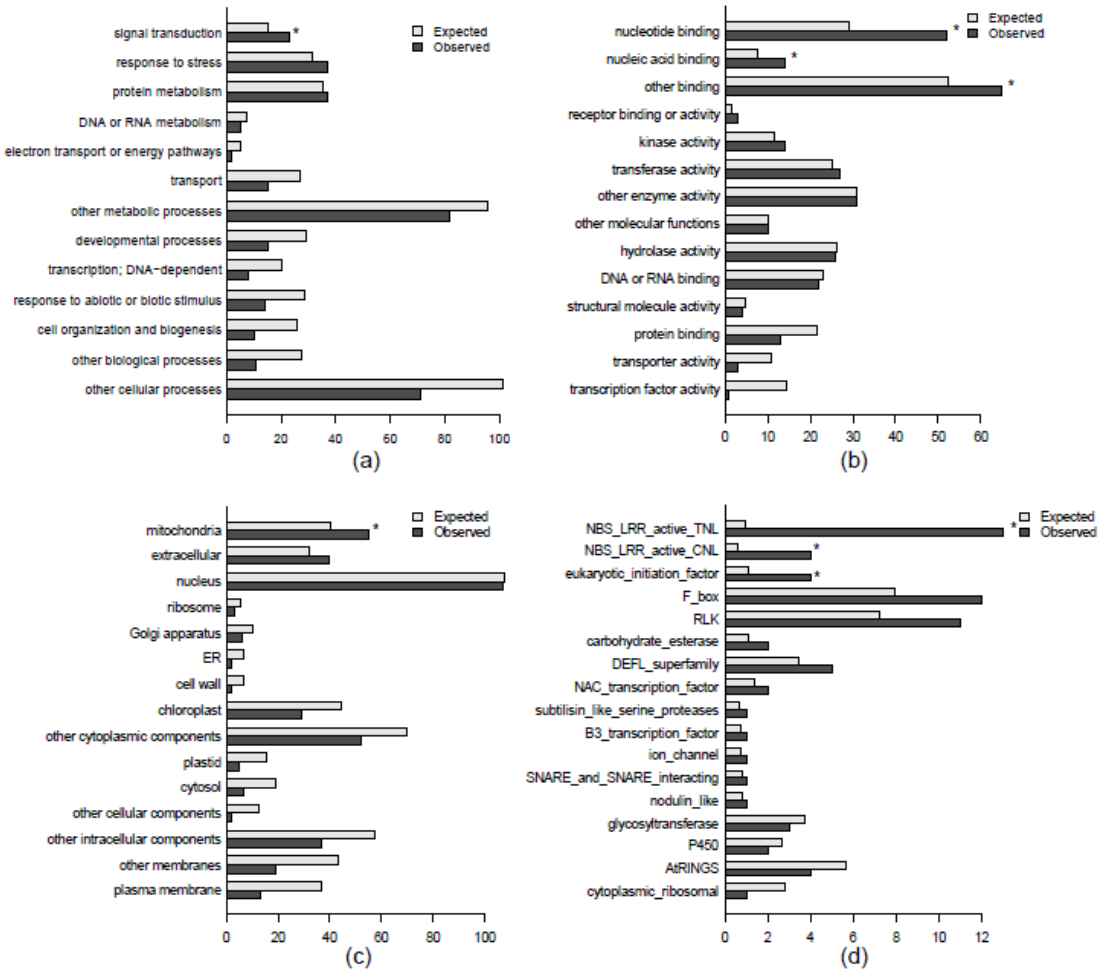


Figure 1. Distribution of 'exon presence/absence' (E-PAV) genes (n=330) – those with at least one, but not all, exons missing in at least one accession – by GOslim categories for molecular function (a), biological process (b) and cellular component (c), and by family (d). Both expected and observed number of E-PAV genes per category represented on each bar. Where there is a significant enrichment (p-value <= 0.05) between the amount of observed and expected E-PAV genes for a particular category an asterisk is shown over the bars. Only categories with at least one E-PAV gene are shown.

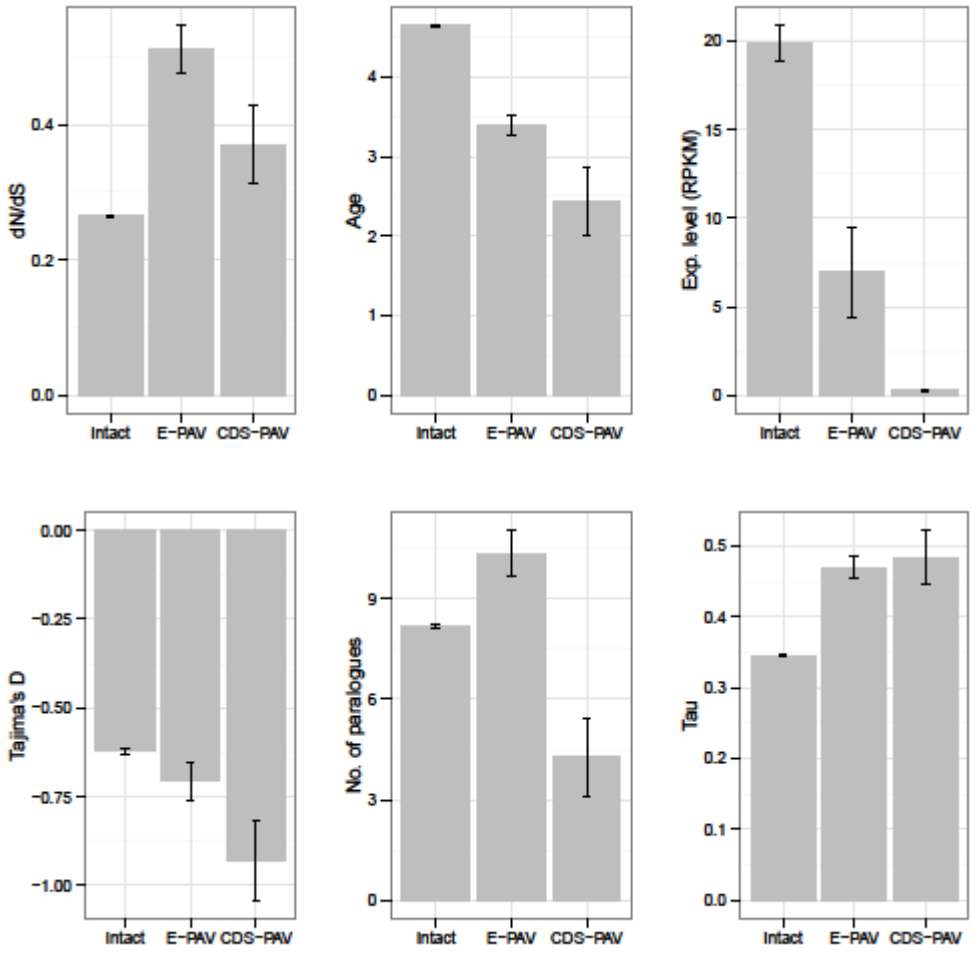


Figure 2. Genetic features associated with intact (having no exons under P/A variation), E-PAV (having at least one, but not all, exons missing in at least one accession), and CDS-PAV (having the entire CDS missing in at least one accession) genes. From left to right, top to bottom: dN/dS, age, expression level, Tajima's *D*, number of paralogues and *tau*. See supplementary table 3 for values of means and statistical analysis.

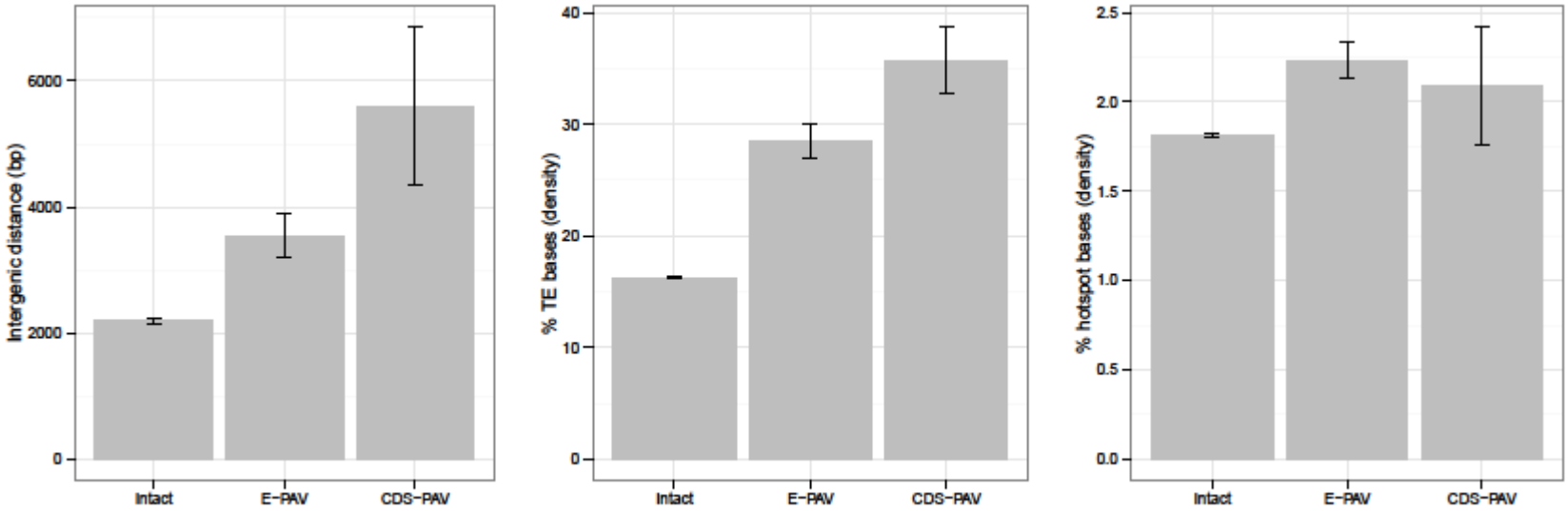


Figure 3. Genomic context for intact (having no exons under P/A variation), E-PAV (having at least one, but not all, exons missing in at least one accession), and CDS-PAV (having the entire CDS missing in at least one accession) genes. Averaged values for the genes in each set are given for, from left to right, the intergenic distance, the percentage of TE bases in the non-genic sequence of a 10kb window centred on that gene's midpoint, and the percentage of recombinogenic motifs in the genic sequence of a 1kb window centred on that gene's midpoint. See also supplementary tables 3 and 4 for the values of specific TE families and other window sizes.

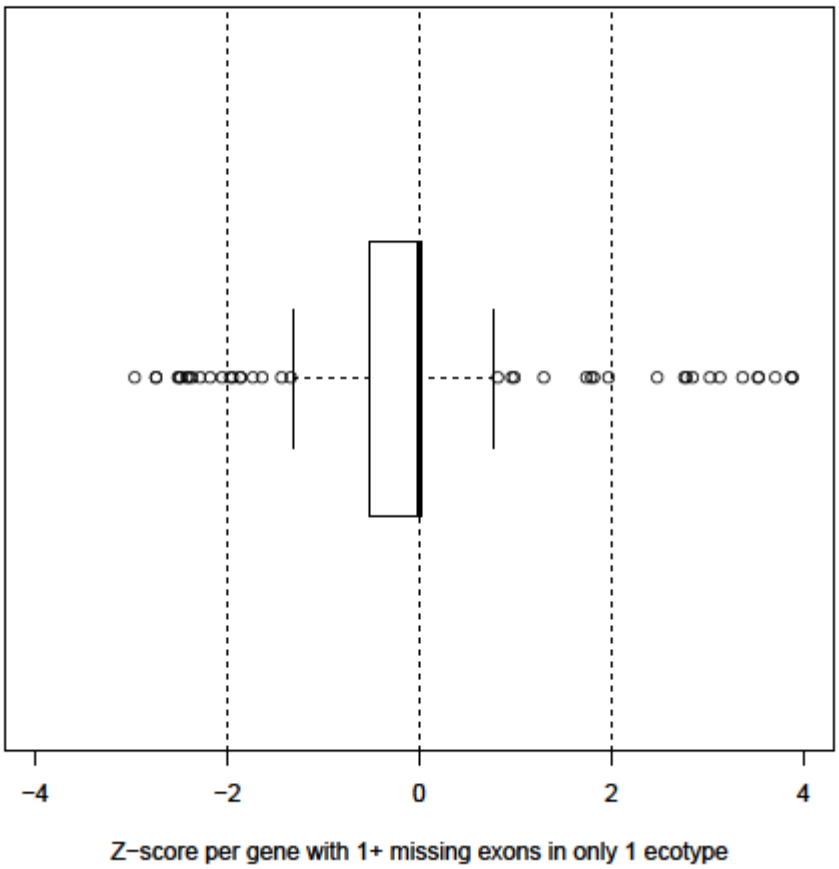


Figure 4. Distribution of Z-scores for standardised transcript abundance data in the affected accession. Data shows 210 genes that have one or more missing exons in only one of 17 *A. thaliana* accessions (relative to Col-0).